This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

THIS PAGE BLANK (USPTO)

CORRECTED VERSION°

PCT

WORLD INTELLECTUAL PRO
International

INTERNATIONAL APPLICATION PUBLISHED UNDE

WO 9603741A1

(51) International Patent Classification 6 : G10L 5/06, 7/08, 9/06, 9/18	A1	(11) International Publication Number: WO 96/03741 (43) International Publication Date: 8 February 1996 (08.02.96)

(21) International Application Number: PCT/US95/09130

(22) International Filing Date: 19 July 1995 (19.07.95)

(30) Pri rity Data:
08/278,266 21 July 1994 (21.07.94) US

(71) Applicant: INTERNATIONAL META SYSTEMS, INC. [US/US]; 6th floor, 100 North Sepulveda Boulevard, El Segundo, CA 90245 (US).

(72) Inventors: PFISTER, Henry, L.; 4455 Torrance Boulevard #147, Torrance, CA 90503 (US). SMITH, George, W.; Suite 200, 23842 Hawthorne Boulevard, Torrance, CA 90505 (US). TSUCHIYA, Masahiro; 1585 Kaminaka Drive, Honolulu, HI 96816 (US).

(74) Agents: STEFFIN, William, C. et al.; Lyon & Lyon, First Interstate World Center, Suite 4700, 633 West Fifth Street, Los Angeles, CA 90071-2066 (US).

(81) Designated States: AM, AT, AU, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IS, JP, KE, KG, KP, KR, KZ, LK, LR, LT, LU, LV, MD, MG, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TT, UA, UZ, VN, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG), ARIPO patent (KE, MW, SD, SZ, UG).

Published
With international search report.

(54) Title: SYSTEM AND METHOD FOR FACILITATING SPEECH TRANSCRIPTION

(57) Abstract

The invention provides a system and method for facilitating speech transcription which accepts continuous speech from any of a variety of conventional devices capable of converting spoken words to electromagnetic signals, including microphones or telephones and, if the input signal is analog, converts the input signal from analog to digital format. The digitized signal is then processed in the time and frequency domains to extract spectral speech features which are used to match the input speech with associated phonemes. According to the invention, possible word choices may be extrapolated from the associated phonemes, and visually displayed in textual representations. The visually displayed text may then be edited and processed into final form. Systems and methods for facilitating the invention are also disclosed.

° (Referred to in PCT Gazette No. 12/1996, Section II)

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgystan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo			SE	Sweden
CH	Switzerland	KR	Republic of Korea	SI	Slovenia
CI	Côte d'Ivoire	KZ	Kazakhstan	SK	Slovakia
CM	Cameroon	LI	Liechtenstein	SN	Senegal
CN	China	LK	Sri Lanka	TD	Chad
CS	Czechoslovakia	LU	Luxembourg	TG	Togo
CZ	Czech Republic	LV	Latvia	TJ	Tajikistan
DE	Germany	MC	Monaco	TT	Trinidad and Tobago
DK	Denmark	MD	Republic of Moldova	UA	Ukraine
ES	Spain	MG	Madagascar	US	United States of America
FI	Finland	ML	Mali	UZ	Uzbekistan
FR	France	MN	Mongolia	VN	Viet Nam
GA	Gabon				

NSDOCID: <WO___9603741A1_IA>

DESCRIPTION

System and Method for Facilitating Speech Transcription

Background of the Invention

The field of the invention is speech recognition and transcription systems. Conventional speech recognition systems have used several independent approaches with limited success. One approach models the vocal tract, articulation, and acoustic production of speech. A second approach models the acoustic waveform and the spectrum of human speech and uses signal processing methods. A third approach models the human ear and its detection mechanisms through neural network methods. A fourth approach analyzes phonetic features, perception, and linguistic models of human speech.

There is a substantial difference in the performance and capability of the systems which currently exist. Initially, these systems accepted only discrete speech since the then nascent technology was unable to distinguish between individual words in continuous speech. This limitation required users to pause between each word. As a result, discrete speech systems proved impractical to use due to the onerous speech adjustments required.

Efforts to improve upon discrete speech systems resulted in the development of systems which could identify one or more designated words within a continuous stream of words. Although these systems could not transcribe continuous speech, they were adequate for applications with limited functionality, such as remote control programming for audio-visual equipment.

A still higher level of complexity achieved by speech recognition systems was the recognition of every word within a sentence. However, these systems required that sentences be constrained within certain grammatical boundaries.

Current efforts in the field of speech recognition have focused on developing speech processors which accept continuous speech while attempting to overcome the above limitations and constraints. These systems perform inadequately when confronted with identifying word boundaries, providing correct spelling, and resolving other word ambiguities.

For example, word boundary issues can arise when any subset of sequential syllables within a multi-syllable word is itself a word. Word boundary problems also arise when the ending syllable or syllables of one word can be joined with subsequent syllables to form another word. Homonyms and multiple spellings of one word also give rise to ambiguities. Therefore, fully automated, accurate end-to-end continuous speech recognition involves intensive data analyses, speech understanding and logical reasoning capabilities. Currently available continuous speech processors thus suffer from transcription errors and other limitations as the requisite technology is currently prohibitively expensive and uneconomical for commercial purposes.

Summary of the Invention

The system and method according to the present invention facilitates speech transcription of normally spoken continuous speech without sacrificing accuracy in transcription. Speech is digitally processed to extract its spectral features. These features are then used to distinguish individual phonemes and to generate a string of equivalent machine-readable phonetic symbols. The string of machine-readable phonetic symbols is processed to identify possible word boundaries for each spoken word within the string. In one embodiment, this process is performed one spoken word at a time. In this embodiment, the possible word choices for each spoken word, as delineated by the identified word boundaries, are visually displayed. The term "word choice" as used herein through

this application represents both singular and multiple word choices as both possibilities may exist and optionally may be displayed for each spoken word. The representation format can vary depending on the particular language used. For example, if the speech is in the Japanese language, word choices may be represented as a string of Hiragana and Katakana symbols, each corresponding to a machine-readable phonetic symbol, or as Kanji characters, or as a combination of Kanji, Hiragana and Katakana letters. Word choices in English, German, French, or other languages may be represented as alphanumeric text or other appropriate representations. In each case, the proper word choice corresponding to each spoken word can be readily selected, thereby ensuring accurate transcription. As the technology develops, it would also be within the scope of the invention to represent the possible word choices in different languages than that dictated. Moreover, the possible word choices can be displayed in various orders, for example, by alphabetical order, by order of probability, or by order of syllabic length, thereby facilitating word selection. Other display orders, including those discussed later, would be apparent to those skilled in the art and are within the scope of the invention.

After the first spoken word within the machine-readable phonetic symbol string is resolved in the above manner, the system proceeds to the next word within the machine-readable phonetic symbol string and iterates the aforementioned process. The process may also use adaptive feedback after each selected word choice to automatically determine the starting word boundary for the next word within the string. Thus, the process always starts at the beginning of the next word, thereby further increasing accurate determination of word boundaries. Moreover, editing tasks such as punctuation, margins, paragraphs, and capitalization, can be performed during this process. Other embodiments process the input speech in larger

blocks of speech. For instance, an embodiment can process the input speech for any block of speech which is bracketed by silence intervals, such as a phrase or sentence at one time. Word choices are then presented for each word within the block of speech at one time.

In contrast to conventional methods of speech recognition which attempt to resolve ambiguities, the invention recognizes that the limitations of currently available conventional technology render this task commercially impractical. Accordingly, in one embodiment, the system according to the invention truncates the conventional process and presents possible solutions to the ambiguities in an environment wherein the correct alternative may be readily selected. Thus, according to the invention, ambiguities may be readily and accurately resolved while repetitive tasks, i.e. collecting, converting and analyzing massive speech data and searching for words in a vocabulary set are performed automatically.

Accordingly, it is an object of the present invention to provide a system and method for facilitating continuous speech transcription. The invention has the further advantages of being both language and speaker independent, in that the invention accommodates any speaker and any spoken language which can be transcribed. Other and further objects and advantages will appear hereinafter.

Brief Description of the Drawings

FIG. 1 is a logical diagram of a preferred embodiment of the invention.

FIG. 2 is a block diagram of speech acceptance and analog-digital hardware suitable for practicing the invention.

FIG. 3 is a block diagram of a spectral processor suitable for practicing the invention.

FIG. 4 is a block diagram of a phoneme labeller suitable for practicing the invention.

FIG. 5 is a block diagram of a speech re-synthesizer suitable for practicing the invention.

FIG. 6 is a graph of the performance of a pre-emphasis filter which is suitable for practicing the invention.

FIG. 7 is a graph and equations showing computation of acoustic power.

FIG. 8 is a graph and equation showing peak-to-peak pitch estimation.

FIGs. 9A-9B are graphs and equations showing frequency domain processing.

FIGs. 10A-10B are an example of forward pass labelling.

FIG. 11 is an example of backward labelling.

Appendix A describes spectral distortion techniques.

Appendix B describes Mel-scale filters.

Appendix C describes weighting techniques.

Appendix D describes forward pass and backward labelling.

Appendix E describes ranking word candidates.

Detailed Description of a Preferred Embodiment of the Invention

While the invention will be described in terms of preferred embodiments, it will be appreciated that other embodiments are possible within the scope of the invention.

As shown in FIG. 1, a preferred embodiment of the system according to the invention has a language, user, and mode selector 2, analog-to-digital convertor 4 inter-acting with an input device 6, a spectral processor 8, phoneme labeller 10 interfacing with phoneme models 14 and preferably, training data 16, a software pre-processor 11, and a word processor 12. In another preferred embodiment, the system may also include a speech re-synthesizer 13.

According to the method of the invention, the speaker dictates speech into the system through the input device

6, which may be a microphone, telephone, other line input, or any input device which converts speech into electromagnetic signals. The term, 'electromagnetic signal', as used throughout this application, includes any form of electrical signals, including microwaves, radio signals, and television signals, as well as optical and other signals such as infrared and laser.

At start-up, the speaker identifies him or herself and selects a language and an operational mode. Three of the possible operational modes are training, dictation and display and editing. These three modes are those known to the inventors at the present time. However, other operational modes may become apparent to one skilled in the art as the technology develops and would also be within the scope of the invention. As explained in detail later, the dictation and display and editing modes may be operated concurrently.

In the training mode, the speaker dictates speech into the system. The system then analyzes the speech to generate data representing the speaker's distinctive voice characteristics. This data is stored as training data and is later used by the system to identify the speaker's words when the system is operated in dictation mode. While not required to practice the broad scope of the invention, completion of the training mode in a preferred embodiment greatly facilitates the efficient practice of the invention.

In dictation mode, the speaker dictates speech into the system, using the input device 6. In one embodiment of the invention, the speaker may dictate the speech at a prior time using conventional means such as a tape recorder and the audio or electromagnetic output of the conventional means can be transmitted to the input device 6. The input speech is passed through the analog-digital converter 4, then through the spectral processor 8, and then through the phoneme labeller 10. The output of the phoneme labeller 10 can either be stored for later

NSDOCID: <WO___9603741A1_IA>

retrieval and processing, or can be sent to the software pre-processor 11. It will be appreciated that elements of the system may be combined or rendered unnecessary as technological capabilities improve, while remaining within the scope of the invention.

In the display and editing mode, editing and other functions may be performed through the software pre-processor 11 and word processor 12, using keyboard, mouse, digitizing pen, voice commands, or any other input-positioning device activated contemporaneously or at a later time by the speaker, by a subsequent user, or by a suitable software/computer combination. The software pre-processor 11 receives the output of the phoneme labeller 10 and works in conjunction with the word processor 12 to display possible word candidates for each spoken word. The displayed word choice is selected by the speaker, a subsequent user, or by an appropriate software/computer combination. The speaker or a subsequent user may also manually input an alternative word choice. This process can then be repeated for the next and subsequent words within the input speech, until the entire input speech has been transcribed. In one preferred embodiment, if a user cannot understand any segment of the spoken input, the system can include a speech re-synthesizer 13 which can reproduce the spoken words. It will be appreciated that within the scope of the invention, the speech input may be stored for later processing. In this embodiment, the speech re-synthesizer can facilitate accurate transcription.

Preferred embodiments of the invention are now described in detail.

<u>Training Mode</u>

Preliminary training is preferably performed to initialize the system for each particular speaker or for each time that a particular speaker uses a new language. During this initialization period, the speaker-specific

training data 16 is established to tune the system to the speaker's individual speech characteristics for the particular chosen language. In this preferred embodiment, the system thus adapts to and accommodates a diverse set of speakers regardless of gender, age, accent, dialect, language or any other factor which can contribute to a difference in the pronunciation of any particular word. Initialization only needs to be performed once for any particular speaker for each particular language.

The speaker-specific training data 16 can be established independent of the invention or a system according to the invention may be used. In embodiments where the speaker generates training data by using the system, the speaker dictates a pre-established reference set of words into the system. Various reference sets of words can be used for any particular language. Methods for determining these reference sets are well-known to those skilled in the art and are also set forth in the publication, Fundamentals of Speech Processing by Rabiner and Juang, Prentice Hall, 1993, which is incorporated herein by reference. The dictated speech is then processed through an analog-to-digital convertor 4, spectral processor 8, and phoneme labeller 10 shown in FIG. 1. The analog-to-digital convertor 4 and spectral processor 8 process the dictated speech in the same manner during a training mode as during a dictation mode and are described in more detail later for the dictation mode.

The phoneme labeller 10 is then used to generate training data 16 by using a spectral distortion measure to compare the speaker's speech against reference speech samples of the pre-determined set of words. Spectral distortion techniques are well-known to those skilled in the art and are also documented in Discrete-Time Processing of Speech Signals by Deller, Proakis, and Hansen, which is incorporated herein by reference and is also attached as Appendix A. Training data can also be

generated independent of the invention through other techniques well known to those skilled in the art.

Preferably, the training data 16 is stored into a data base. The data base may be organized in a variety of suitable structures; however, a preferred embodiment uses the same structure as in DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM NISTIR 4930, NTIS, February, 1993, which structure is shown in Example 1:

Example 1:

database hierarchy:
/<corpus>/<usage>/<dialect>/<sex>/<speaker>/<word id>.file
corpus:= language
usage:= train/test
dialect:=$d_1 \ldots d_n$
sex:= m/f
speaker:= (initials)(digit)
word id:= <text type><word number>
text type:= sa/si/sx
word number:= 1...n words
file:= wav/txt/wrd/phn

Dictation Mode

Speech Acceptance and Analog-Digital conversion.

In ordinary use, that is, after the system has received and processed training data for the speaker and the chosen language, the system of this preferred embodiment is used in dictation mode. As indicated in FIG. 2, speech is received into the system via any of a variety of conventional devices, including but not limited to static or dynamic microphones 18, telephones 22, or other line input 20. As will be appreciated by those skilled in the art, speech input could also be transmitted and received by other technology via electromagnetic signals.

If microphones are used, they preferably have noise limiting capability. As shown in FIG. 2, appropriate amplifiers 24 and interfaces 26 are used to connect the

SDOCID: <WO___9603741A1_IA>

receiving devices with a low-pass analog filter 28 and to ensure a fixed level output compatible with digital equipment. These devices are readily available. Moreover, one skilled in the art would know how to choose appropriate amplifiers and interfaces. The text, Digital Signal Analysis, by Stearns, Hayden 1975 can also be referenced to determine the compatibility of these devices.

Once received, the signal representing the input speech is passed through a low pass analog filter 28 which eliminates frequencies that can alias into the acoustic signal during subsequent processing. One preferred embodiment uses a 3rd order Butterworth analog filter with 10 kHz and 6 dB/octave as the low-pass filter of choice.

The signal is then passed through an analog-to-digital converter 30. While a variety of analog-to-digital converters may be used for this purpose, a preferred embodiment uses a Motorola 56 ADC 16 with sixteen bit resolution at 20 Khz and a continuously adjustable gain over a 20 db range.

Phonetic Feature Detection

The digitized signal is passed through a spectral processor, as shown in FIG. 3, to extract spectral speech features. Before describing this process in detail, it is useful to discuss the problem of environmental or "background" noise which is commonly encountered in speech recognition. Two classes of problems contribute to background noise. One class is attributed to the speaker and consists of sound artifacts such as lip smacks, heavy breathing, mouth clicks, and nasal pops. Although such artifacts are generated inadvertently, they often have an energy level comparable to speech. As will be further discussed, the invention preferably models speaker artifacts and detects them along with phoneme recognition so that they may be removed during subsequent processing.

A second class of noise problems arises from the ambient noise environment. Occasionally, non-speaker

generated background noise, such as a bell, whistle or other sound, may interfere with and mask the speech. To alleviate this problem, the system can use a noise limiting microphone for receiving the dictated speech. Other means of eliminating background noise will be apparent to those skilled in the art.

Additionally, or independently, a running average of the ambient background noise during periods of silence can be kept so that sudden non-vocal tract background noises can be easily classified and eliminated. As will be explained in more detail, a preferred embodiment uses both a noise limiting microphone and continuous modelling of background noise. These techniques avoid the problems commonly experienced by other speech processing systems that pre-process the signal to remove noise, as the latter lose or distort the speech.

In a preferred embodiment, the digitized signal is first passed through a pre-emphasis filter 32 which works in connection with a noise model 34 to eliminate background noise. The noise model 34 is initially set to a default noise level and is adaptively updated during subsequent processing of the input speech. This adaptive feature is later discussed in connection with frequency domain processing.

In one preferred embodiment, the pre-emphasis filter 32 is a Finite Impulse Response (FIR) bandpass digital filter with zero phase shift and unity gain. The derivation of the bandpass coefficients for such a filter is shown in Example 2.

<u>Example 2.</u>

A 20 khz sampling rate with a bandpass from 100 hz to 6000 hz is assumed. The sample signal has a 60 hz hum component, a 7000 hz whistle, random noise, and an acoustic signal at 1000 hz.

Filter design characteristics:

Lower cutoff frequency: wlo = 100

Upper cutoff frequency: whi = 6000
Sample period: dt = .00005
rate: = 1/dt = 2 X 10^4
Number of coefficients: nc = 45

Filter coefficients computation:

whir = whi * 2 * π = 3.77 X 10^4
wlor = wlo * 2 * π = 628.319
c_0 = dt * (whir - wlor) / π = 1.8574
i = 1 .. nc; c_i= (sin(i*whir*dt) - sin(i*wlor*dt))/(π*i)

Sample signal:

nn = 511 nt = 2 noise = 0.5
f_0 = 60 f_1 = 1000 f_2= 7000
a_0 = 0.5 a_1= 1.0 a_2= 0.5
i = 0 .. nn

$$input_i=\sum_{j=0}^{nt} a_j*\sin(i*dt*2.0*\pi*f_j)+rnd(noise)-\frac{noise}{2}$$

$output_i$ = 0
$error_i$ = 0
$signal_i$ = sin(i*dt*2.0*π*f1)
k= nc..nn-nc

$$output_k=c_0*input_k+\left[\sum_{j=1}^{nc}[(c_j*input_{k+j})+(c_j*input_{k-j})]\right]$$

$$error_k=|output_k-signal_k|$$

The performance of the pre-emphasis filter 32 is shown in FIG. 6. In FIG. 6, the upper trace (a) is the acoustic input signal. The middle trace (b) is the result of the pre-emphasis filter. The third trace (c) is the

NSDOCID: <WO___9603741A1_IA>

desired signal that is embedded in the signal. The lower trace (d) is the absolute difference of the acoustic signal and the filter output.

The signal is then sampled into blocks of data as indicated by the sampler 36. Since the human vocal mechanism modulates the slowly changing speech signal onto a higher frequency sound wave, it would be necessary to sample the acoustic wave form at over 10,000 times per second to capture this speech wave. In contrast, the movement of the tongue, jaw, lips and other vocal articulators change at the far slower rate of less than 100 times per second. This physical situation is exploited by grouping acoustic data into approximately one hundredth of a second blocks to isolate phonetic features and to identify noise sources.

Taking the above considerations into account, a preferred embodiment samples the signal into blocks of 512 samples. Oversampling can be used to increase accuracy during subsequent processing. Thus, a preferred embodiment uses a 25%, or 128-sample overlap between adjacent blocks. As can be readily appreciated by one skilled in the art, numerous other data block sizes and overlap ranges can be used. For example, blocks of 256, 1024 or 2048 samples may be used. Moreover, any overlap range between 0 and 50% may be used between adjacent blocks.

While sampling may be accomplished with hardware, microcode, or other methods known to those skilled in the art, a preferred embodiment uses microcode. As technology advances, other sampling methods and devices suitable for practicing the invention may become available.

The sampled signal is then processed in the time domain, as indicated by element 38, to extract the spectral speech features of acoustic power and the peak-to-peak pitch.

Acoustic power indicates the presence of speech or silence intervals that naturally occur at the end of sentences, phrases, or words. To determine acoustic

power, a short time energy estimate is made for each sample block by averaging either the magnitude or square of the signal within the block, otherwise referred to as absolute power and squared power, respectively. Next, the peak-to-peak pitch is estimated by summing the signal zero crossings. Although this process is well-known to one skilled in the art, an example of time domain processing is given in Figures 7-8.

In FIG. 7, a sample block of 512 samples is shown with a mean value of $3.298 * 10^4$. Plot (a) represents the sample block signal. Equations (b) and (c) are used to compute the absolute power and squared power, respectively.

In FIG. 8, the peak-to-peak pitch is estimated by summing the signal zero crossings of the estimated signal shown in plot (d), as shown by equation (e).

The signal is also passed through acoustic band pass filters 40 to identify the relative signal power in each acoustic band, otherwise referred to as the spectral pattern. This information can be used later to assist in phoneme identification. In a preferred embodiment, the band pass filters are selected based on the Mel scale. The Mel scale is a logarithmic-based scale which is more fully described in the publication <u>Advanced Algorithms and Architectures for Speech Understanding</u>, ESPRIT Research Report Project 26, Vol. 1, Pirani, 1990, and is herein incorporated by reference, and is also attached as Appendix B.

The estimated formant frequencies are determined by processing the signal in the frequency domain as shown by element 42. An example of frequency domain processing is given in Figures 9A-9B. In Figure 9A, plot (a) represent the same sample block as plot (a) in Figure 7. The sample block can be first tapered using a tapering function. Numerous tapering functions, including a Hanning Window as shown in Equation (b), can be used to taper the 512 elements of the sample block, and are well-known to those

skilled in the art. Tapering reduces high frequency noise in the subsequent Fast Fourier Transform (FFT). Since the signal was previously oversampled by 256 elements (128 elements at each end), tapering promotes accuracy, rather than loss of data. Plot (c) represents the tapered sample block. As would be appreciated by one skilled in the art, numerous other tapering functions can be used.

An FFT is then performed on the tapered sample block and the power spectrum is estimated using the square of the resulting complex frequency coefficients, as shown by equation (d) in FIG 9B. Plot (e) provides a graph of the resultant power spectrum. The log of each element is then computed, as shown in plot (f). Since human speech usually does not exist at frequencies greater than 8000 hz, the frequency spectrum in plot (f) can be truncated to reduce high frequency noise. The cut-off range can be as low as 7500 hz or as high as 8500 hz. In the embodiment shown in plot (g), the frequency is truncated at 7500 hz. Next, an inverse FFT is performed to determine the cepstral coefficients, as shown in plot (h). The resonant or formant frequencies of the vocal tract are then extracted from the low order peaks of the cepstrum in plot (h).

Note that if a silence interval is detected during time domain processing, the data from frequency domain processing necessarily represents only ambient background noise. The system and method according to the invention update the noise model 34 with this data. Thus, the noise model 34 reflects a semi-continuous model of background ambient noise. This adaptive feedback enhances subsequent processing of speech through the pre-emphasis filter 32.

Thus, in this embodiment, the spectral processor extracts the following spectral speech features: acoustic power, spectral pattern, and formant frequencies.

Phoneme Identification.

The extracted spectral speech features are passed to a phoneme labeller 10 as shown in FIG. 4. As would be

known to one skilled in the art, the previously extracted spectral feature of relative signal power can be used to classify phonemes into one of the following phonetic categories: vowels, diphthongs, liquid or glide semi-vowels, nasal consonants, voiced or unvoiced fricatives, africates, whisper and voiced or unvoiced stops. Table 1 describes a possible phoneme classification scheme for forty standard phonemes used in Western/European languages:

<u>Table 1:</u>

- vowels divided into:
  - front -- IY, IH, EH, AE
  - mid -- AA, ER, AH, AX, AO
  - back -- UW, UH, OW
- dipthongs - AY, OY, AW, EY
- semivowels, including
  - liquids -- WW, LL
  - glides -- RR, YY
- nasal consonants -- MM, NN, NG
- stops
  - voiced -- BB, DD, GG
  - unvoiced -- PP, TT, KK
- fricatives
  - voiced -- VV, TH, ZZ, ZH
  - unvoiced -- FF, TH, SS, SH
- africates -- JH, CH
- whisper -- HH

By classifying phonemes into phonetic categories, the individual phoneme identification problem can be greatly simplified. For instance, under the classification scheme given in Table 1, each phonetic category contains at most five elements out of a possible forty. Moreover, these elements are usually well separated by physical features of the vocal tract so that subsequent identification of an individual phoneme within a phonetic category is enhanced.

To identify individual phonemes for each signal block, a 'forward pass' 44 is made through each signal block and phoneme candidates are identified by comparing against existing features in the phoneme models 14. The phoneme models 14 are pre-determined reference sets of data for each language. Techniques for generating these sets are well-known. For instance, the TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM NISTIR 4930, NTIS, February, 1993 is one such set which is readily available to one skilled in the art. A score is kept for each phoneme candidate based on its probability of being the correct phoneme spoken.

In a preferred embodiment, scoring is accomplished by using the following fuzzy logic LR membership functions:

Left(z) = 1/(1 + z*z) Left side function

Right(z) = 1/(1 + z*z) Right side function

The function is described as a four point trapezoid (vw) and disjunction:

$$ww(a,b,c,d,y,x) = if\left(x<b, left\left(\frac{b-x}{b-a}\right), if\left(x>c, right\left(\frac{x-c}{d-c}\right), y\right)\right)$$

mx(a,b) = if(a>b,a,b) mn(a,b) = if(a<b,a,b)

wx(a1,b1,c1,d1,a2,b2,c2,d2,y,x) = ww(mx(a1,a2),mx(b1,b2),mn(c1,c2),mn(d1,d2),y,x)

These functions are well-known to those skilled in the art and are also discussed in <u>Fuzzy Set Theory and Its Applications</u> by Zimmermann, 2d, Kluwer, 1991, which is herein incorporated by reference. In this preferred embodiment, each phoneme candidate feature and sample feature are used to form a normalized membership function of the sample in the phoneme data base. Thus, all phonemes have a possibility of assignment in each sample block. The features can be weighted to maximize the probability of phoneme identification. The weighting

process can be accomplished through a variety of techniques which are well-known to those skilled in the art, some of which are described in Discrete-Time Processing of Speech Signals, which are incorporated herein by reference and are also attached as Appendix C. Preferably, the determination of the weighting function is based on experimental scoring derived from the training data 16 for the speaker. As discussed later, the weighting function can be updated through adaptive feedback from the user. An example of forward pass labelling is given in FIGs. 10A-10B.

As previously mentioned, speaker artifacts and sudden, loud background noises can interfere with the input speech so that the system is unable to find a reasonable phoneme match for the received sound. In a preferred embodiment, such interference is identified and the speaker is requested to repeat the masked speech.

When the extracted spectral speech features indicate a silence interval, a "speech segment" bracketed by the silence interval and its initial starting point has been defined. This event triggers backward labelling 46 of the most likely phoneme candidates for the speech segment. Simultaneously, spectral processing 8 and forward pass labelling 44 continue for the next and subsequent speech segments within the speech signal.

A silence interval could be defined as any time interval longer than 1/10 (one-tenth of a second) in which the ambient noise level does not exceed 10 dB over the average ambient background noise level. Shorter time intervals and smaller decibel increases may also be used. For instance, a shorter time period for speakers who dictate too rapidly to pause for more than 1/10 of a second, or for speakers who speak too softly to develop a 10 dB increase over the ambient background noise level. Thus, it is also within the broad scope of the invention for embodiments to define silence intervals as short as 1/15 of second, or to require only a 9 dB increase. Other

embodiments may require over 12 dB increases for speakers who are particularly loud. Other combinations are readily apparent to those skilled in the art.

Backward labelling for these subsequent speech segments is initiated once processing for the instant speech segment is complete. Notably, partitioning of the speech into speech segments bracketed by silence intervals promotes accuracy and facilitates parallel processing. Moreover, bracketing further facilitates the display of full phrases in embodiments which process speech in blocks of phrases at one time. These embodiments are described later.

Backward labelling identifies the N most possible phoneme sequences within the speech segment. Each phoneme candidate is ranked by its possibility of assignment in each sample block, based on conventionally known methods of phoneme identification. These methods are discussed in Discrete-Time Processing of Speech Signals, and are incorporated herein by reference, and are also attached as Appendix D. As shown in FIG. 11, the candidates are sorted to maximize the possibility of phoneme identification.

Once the phoneme candidates are identified, the system generates an equivalent machine-readable phonetic symbol for each phoneme candidate, as shown by element 48. This process utilizes look-up tables 15 which associate machine-readable phonetic symbols with each phoneme model 14. The machine-readable phonetic symbols are composed of machine-readable codes. Thus, by cross-referencing each phoneme candidate with the equivalent phoneme model and its associated machine-readable phonetic symbol, a corresponding machine-readable phonetic symbol is generated for each phoneme candidate. Although various machine-readable codes can be used, a preferred embodiment uses ASCII codes.

Output to Software Pre-Processor

In one preferred embodiment, the machine-readable phonetic symbols are directly output to a software pre-processor 11 as shown in FIG. 1. The software pre-processor 11 can be separately joined to or integrated with a word processor 12 running, for example, on a personal computer, on stand-alone equipment or as part of a network or central computer system. The software pre-processor 11 may be connected to the phoneme labeller 10 via cable connection or any conventionally available wireless communication device such as infrared transmittal, laser, radio, microwaves, or television. The process then continues with the editing mode described later.

In another preferred embodiment, the machine-readable phonetic symbols can be stored on any conventionally available device which stores machine-readable data, such as, for example, computer data disks, hard drives, flash, static or dynamic memory, tape, or CD-PROM. Once the machine-readable phonetic symbols are stored, the process can be re-started at a later time by loading the machine-readable phonetic symbols into the software pre-processor 11 and continuing with the display and editing mode described below.

Another preferred embodiment automatically stores the machine-readable phonetic symbols while outputting the machine-readable phonetic symbols to the software pre-processor 11.

Display and Editing Mode

Once the software pre-processor 11 starts to receive the machine-readable phonetic symbols corresponding to the last spoken speech, the system can be operated in the display and editing mode, utilizing the software pre-processor 11, word processor 12, and speech re-synthesizer 13, as described below. In one embodiment, the display and editing mode occurs after the speaker has finished

dictating, as the machine-readable phonetic symbols which represent the dictated speech can be stored for later processing in the display and editing mode. However, it is important to note that the speaker may continue dictating while words spoken earlier are being processed through the display and editing mode. For example, a secretary may operate the display and editing mode while the speaker continues to dictate. Alternatively, the speaker may pause during dictation to edit prior dictation.

Phonetic Symbol String Editing

In a preferred embodiment, the software pre-processor 11 provides an option to edit the machine-readable phonetic symbol string. This option is particularly useful for written languages which can be readily comprehensible as phonetic symbols, such as Japanese Hiragana and Katakana. To edit the machine-readable phonetic symbol string, the software pre-processor 11 outputs the machine-readable phonetic symbol string to the word processor 12, where it is displayed.

If a wrong machine-readable phonetic symbol or set of symbols is detected by the speaker or subsequent user, the symbol or symbols can be manually overridden by keyboard, mouse, digitizing pen, voice command or any other input-positioning device such as a track ball, joystick or touchscreen. For voice-input, the user manually selects the subject symbol or symbols using either a keyboard, mouse, digitizing pen, or other input-positioning device, and then re-dictates the input speech as desired. This speech is then processed as before, i.e. through the analog-digital convertor 4, spectral processor 8, phoneme labeller 10 and software pre-processor 11 to update the machine-readable phonetic symbol string.

If the input speech generates more than one possible machine-readable phonetic symbol match during backward labelling 46, each possibility can be displayed so that the speaker or subsequent user can readily select the

correct machine-readable phonetic symbol. A preferred embodiment only displays the two most likely machine-readable phonetic symbols.

Thus, the machine-readable phonetic symbol string may be edited using the software pre-processor 11 and word processor 12. Certain languages, such as English, German, and French, differ from Japanese in that no corresponding syllabic symbols, such as Hiragana and Katakana, exist for the former. In these cases, a universal phonetic symbol representation can be used. Alternatively, the speaker or subsequent user can opt to skip machine-readable phonetic symbol string editing altogether and proceed directly to locating word boundaries.

Identifying word boundaries

The software pre-processor 11 identifies word boundaries within the machine-readable phonetic symbol string. Each substring of machine-readable phonetic symbols delineated by the identified word boundaries is combined into a possible word choice, which is represented in machine-readable code. While this process can be performed for the entire machine-readable phonetic symbol string or any segment thereof at one time, a preferred embodiment identifies word boundaries on a word-by-word basis. In this preferred embodiment, the software pre-processor 11 starts at the beginning of the machine-readable phonetic symbol string and identifies possible word boundaries within the string for the first spoken word. Identification of possible word boundaries for subsequent spoken words occurs later, and can be accomplished, in part, by using adaptive feedback to determine the next starting word boundary. This embodiment is described in detail later.

As an example of identifying word boundaries for Japanese speech, the software pre-processor 11 identifies every Kanji character which can be represented by any substring of machine-readable phonetic symbols with a

beginning word boundary at the start of the machine-readable phonetic symbol string. These Kanji characters are each considered a possible word choice. Techniques for combining machine-readable phonetic symbols into words are commonly known. For instance, commercially available Japanese word processors operate on these principles to convert Hiragana and Katakana symbols into Kanji characters.

The possible word choices delineated by the identified word boundaries are then ranked in order of probability by reference to linguistic usage, including such factors as grammatical and contextual syntax. Techniques for performing this task are well-known to those skilled in the art and are also described in the aforementioned text, Discrete-Time Processing of Speech Signals, which are herein incorporated by reference and are also attached as Appendix E. Alternatively, the possible word choices may be ranked by alphabetical or reverse alphabetical order, or by increasing or decreasing syllable length, where syllable length is measured by the number of syllables within a word.

An alternative method of ranking the word choices is with reference to the prior usage of words in the dictated speech. In this method, previously dictated and selected words, as later described, are accessed and if matches are found, the word choices presented reflect the previous usage. The word choices can then be ranked, for example, by their most recent usage, by the time elapsed between the current words and their prior usage (i.e. age), or by frequency of usage.

The possible word choices are then output, in machine-readable form, to the word processor 12, or are stored for future use.

Visually Representing Possible Word Choices

The possible word choices are then visually represented. In a preferred embodiment, the word processor 12

SDOCID: <WO__9603741A1_IA>

possesses all the features commonly available on commercial word processors. The word processor 12 displays the possible word choices in representations readily comprehensible to the speaker or a subsequent user. The representation format may vary depending on the language used. For instance, in Japanese, the word processor can represent each possible word choice as a Kanji character, Hiragana and Katakana representations, or a combination of the three. For English, German, French and other similar languages, the possible word choices can be represented as alphanumeric text or other appropriate representations.

Numerous display orders are also possible. For instance, the possible word choices can be displayed in order of probability, alphabetical order, reverse alphabetical order, by increasing or decreasing syllable length, where syllable length is measured by the number of syllables in a word, by the most recent usage, by the time elapsed between the current word and the word choice's prior usage (i.e. age), or by frequency of usage.

In a preferred embodiment, the most probable word choice is displayed separate from the remaining word choices and the remaining possible word choices are then displayed as two independent sequences. One of these sequences ranks the remaining word choices by increasing syllable length and the other sequence is ranked by alphabetical order.

There are numerous methods by which the most probable word choice can be displayed separate from the remaining words. For instance, the most probable word choice can be displayed in the first line and the two independent sequences of remaining words as separate columns below the first line. Alternatively, the most probable word choice can be displayed in **bold** format as compared to the remaining words. Other variations are readily apparent to one skilled in the art.

Word Selection

The word processor 12 further provides the speaker, a subsequent user, or a suitable software/computer combination with functionality to readily select any displayed word choice or to manually input a word. These options may be similar to spell checking features commonly available on conventional word processing software packages. The selection process may be performed by any one of numerous means, including but not limited to voice command, mouse, digitizing pen, keyboard or any other input-positioning device. Insertion of punctuation and similar tasks can be accomplished at this time either manually, or through macros or voice commands.

A preferred embodiment further provides an option whereby the speaker or a subsequent user may call up the machine-readable phonetic symbol substring associated with any displayed word. The string may then be edited in the same manner as before, after which the software pre-processor 11 re-translates the modified machine-readable phonetic symbol substring into a word and displays the word for further editing. Alternatively, the user may manually select word boundaries within the recalled machine-readable phonetic symbol substring, which the software pre-processor 11 then uses to re-combine the machine-readable phonetic symbols into words which are displayed for acceptance or further editing.

It is recognized that someone other than the original speaker (referred to herein as "subsequent user") may use the system to actually perform the editing process. Thus, a preferred embodiment of the system has the additional capability of re-synthesizing the sound represented by any speech segment so that the user can hear what speech was actually dictated.

As shown in FIG. 5, the speech re-synthesizer first uses a digital signal generator 50 which receives machine-readable phonetic symbols from the software pre-processor 11 and utilizes the look-up tables 15, and if desirable,

training data 16, to convert the machine-readable phonetic symbols into a digital signal. This digital signal is fed to a digital-to-analog convertor 52, and then to a speaker driver 54 connected to an audio speaker 56. Each of these devices is commercially available.

Accordingly, the speaker need not personally perform the editing process since any ambiguities can be understood by any subsequent user, especially in embodiments incorporating re-synthesized speech processes. It is recognized that this particular re-synthesis feature can be implemented for uses other than speech transcription. For instance, storing speech as an analog signal consumes a relatively large amount of storage media. Even a digitized speech signal requires a substantial amount of storage space. However, storing speech as machine-readable phonetic symbols requires relatively little space. Thus, the above speech re-synthesis technique constitutes an ideal speech compression method which is also readily adapted to telephone answering machines and other voice-storage applications.

Adaptive Feedback

In a preferred embodiment, the word processor 12 utilizes adaptive feedback so that information as to a selected word is incorporated into the selection of the next starting word boundary. After the correct word choice has been selected from the possibilities presented, this information is fed back to the software pre-processor 11 so that the software pre-processor will begin at the next machine-readable phonetic symbol. Thus, subsequent word boundary identification is enhanced since the beginning word boundary has already been determined.

In addition to resolving word boundaries, each selection of an ambiguous word necessarily resolves ambiguities in phoneme identification. A preferred embodiment exploits this situation by also using adaptive feedback so that user selection information can be used to

update the weighting functions for subsequent phoneme labelling. Thus, this feature allows the system and method according to the invention to further adapt to a speaker's particular speech characteristics.

It is recognized that this feedback capability may not always be desirable. For instance, a speaker's speech characteristics may vary from time to time, due to general health, stress or other factors. Feedback may be undesirable in these instances, since the weighting functions will be updated with anomalous data. Therefore, unlike feedback for determining starting word boundaries, an option is provided whereby feedback to the weighting functions can be disabled.

In practice, feedback to the weighting functions is preferably used only for a short period of time, depending on the consistency of an individual speaker's speech. After the weighting functions have been updated during this period, feedback to the weighting function is disabled. Feedback may be subsequently enabled to update the weighting functions if a speaker's speech characteristics have changed over a period of time, for instance, as a speaker grows older.

<u>Subsequent Word Resolution</u>

Once the first word is resolved in the above manner, the process is iterated for the next and subsequent words. As previously discussed, once the first word has been selected, the software pre-processor 11 begins at the next machine-readable phonetic symbol within the machine-readable phonetic symbol string to identify possible word boundaries for the next word.

<u>Other Embodiments for Longer Speech Blocks</u>

While a preferred embodiment analyzes the machine-readable phonetic symbol string on a word-by-word basis, it is within the scope of the invention to analyze larger blocks of speech at one time, for instance, sentences or

phrases, or any blocks of speech which are bracketed by silence intervals.

These embodiments use the same principles as in a preferred embodiment. For instance, if the speech is processed in a phrase or sentence block, the software pre-processor 11 determines word boundaries for all the words in the entire phrase or sentence, which is then displayed through the word processor 12. The user is then given the option of validating the entire phrase or sentence as correct, or manually editing any word boundary within the displayed block. In the latter case, the software pre-processor 11 re-identifies word boundaries for the remaining words within the displayed block, since the user's editing may change subsequent word boundaries within the sentence. The updated block is then re-displayed for acceptance or further editing.

Accordingly, a system and method for facilitating speech transcription has been disclosed. Although preferred embodiments are described to process continuous speech, it will be appreciated that the scope of the invention will also accommodate discrete speech. It is further apparent that a preferred embodiment may be implemented in any spoken language which can be transcribed and is also speaker-independent. Many other embodiments are easily recognized by one skilled in the art and are within the scope of the invention. For instance, the invention may be implemented to process speech on a phrase-by-phrase or sentence-by-sentence basis. A person skilled in the art could readily modify the system to accept digital input, as from an audio stereo system, by omitting analog-to-digital conversion and making appropriate modifications to pre-process the digital signal. Although currently not commercially feasible, another embodiment can use a computer, such as a Sun 10 Workstation, with sufficient computing capability to perform the editing and selecting process. The

invention, therefore, is not to be restricted except in the spirit of the appended claims.

Claims

1. A system for facilitating speech transcription comprising:

(a) a device which receives a signal representing speech;

(b) an analog-to-digital convertor in communication with said device for converting said signal into a digital signal;

(c) a spectral processor in communication with said analog-digital convertor for receiving said digital signal and extracting spectral speech features from said digital signal;

(d) a phoneme labeller in communication with said spectral processor for receiving said spectral speech features, identifying phonemes from said spectral speech features and generating corresponding machine-readable phonetic symbols for said phonemes;

(e) a software pre-processor in communication with said phoneme labeller for receiving said machine-readable phonetic symbols as input and for combining said machine-readable phonetic symbols into words; and

(f) a word processor in communication with said software pre-processor for visually displaying said words.

2. The system as in claim 1 wherein said signal representing speech is an electromagnetic analog signal.

3. A method for facilitating the transcription of speech comprising the steps of:

(a) receiving a digital signal representing human speech;

(b) processing said digital signal to extract spectral speech features;

(c) identifying phonemes from said spectral speech features;

(d) generating corresponding machine-readable phonetic symbols for said phonemes;

(e) outputting said machine-readable phonetic symbols into a software pre-processor;

(f) identifying word boundaries within a substring of said machine-readable phonetic symbols;

(g) combining said substring into a possible word choice; and

(h) visually representing said possible word choice.

4. A method for facilitating speech transcription comprising the steps of:

(a) receiving a human voice speech input as an analog signal;

(b) converting said speech input from said analog signal to a digital signal;

(c) processing said digital signal to extract spectral speech features;

(d) identifying phonemes from said spectral speech features;

(e) generating corresponding machine-readable phonetic symbols for said phonemes;

(f) outputting said machine-readable phonetic symbols into a software pre-processor;

(g) identifying word boundaries within a substring of said achine-readable phonetic symbols;

(h) combining said substring into a possible word choice; and

(i) visually representing said possible word choice.

5. The method as in claim 4 wherein said step (d) of identifying phonemes comprises comparing said spectral speech features with a reference set of phoneme models.

6. The method as in claim 5 wherein said step (d) of identifying phonemes further comprises comparing said spectral speech features with speaker-specific training data generated prior to said step (d).

7. The method as in claim 4, 5, or 6 wherein said step (i) of visually representing said possible word choice comprises displaying said possible word choice as alphanumeric text.

8. The method as in claim 4, 5, or 6 wherein said step (i) of visually representing said possible word choice comprises displaying said possible word choice as every machine-readable phonetic symbol within said substring.

9. The method as in claim 4, 5, or 6 wherein said step (a) of receiving a human voice speech input comprises using at least one of a static microphone, a dynamic microphone, a telephone, and line input.

10. The method as in claim 4, 5, or 6 wherein said step (a) of receiving a human voice speech input occurs in a selected language and said steps (d) - (h) and said step (i) of visually representing said possible word choice occur in said selected language.

11. The method as in claim 4, 5 or 6 further comprising the step of ranking said possible word choice by increasing syllable length, wherein the word choice with the fewest number of syllables is ranked first.

12. The method as in claim 4 further comprising the step of ranking said possible word choice by decreasing syllable length, wherein the word choice with the most number of syllables is ranked first.

13. The method as in claim 4 further comprising the step of ranking said possible word choice by alphabetical order.

SDOCID: <WO___9603741A1_IA>

14. The method as in claim 4 further comprising the step of ranking said possible word choice by probability in reference to linguistic usage.

15. The method as in claim 14 wherein said step (i) of visually representing said possible word choice comprises displaying the most probable word choice in reference to linguistic usage.

16. The method as in claim 4 further comprising the step of ranking said possible word choice by at least one of alphabetical order, reverse alphabetical order, probability in reference to linguistic usage, increasing syllable length and decreasing syllable length.

17. The method as in claim 16 wherein said step (i) of visually representing said possible word choice comprises displaying the most probable word choice, in reference to linguistic usage, separated in the display from the other choices and displaying the remaining word choices in alphabetical order.

18. The method as in claim 16 wherein said step (i) of visually representing said possible word choice comprises displaying the most probable word choice, in reference to linguistic usage, separated in the display from the other choices and displaying the remaining word choices in reverse alphabetical order.

19. The method as in claim 16 wherein said step (i) of visually representing said possible word choice comprises displaying the most probable word choice, in reference to linguistic usage, separated in the display from the other choices and displaying the remaining word choices in increasing syllable length.

DOCID: <WO__9603741A1_IA>

20. The method as in claim 16 wherein said step (i) of visually representing said possible word choice comprises displaying the most probable word choice, in reference to linguistic usage, separated in the display from the other choices and displaying the remaining word choices in decreasing syllable length.

21. The method as in claim 16 wherein said step (i) of visually representing said possible word choice comprises displaying the most probable word choice, in reference to linguistic usage, separated in the display from the other choices and displaying the remaining word choices in order of probability in reference to linguistic usage.

22. The method as in claim 4, 5 or 6 wherein said step (c) of processing said digital signal comprises sampling said digital signal into blocks of samples.

23. The method as in claim 22 wherein said blocks of samples have a 25% overlap between adjacent blocks.

24. The method as in claim 22 wherein said step (c) of processing said digital signal comprises Fast Fourier Transform after first tapering each block of samples.

25. The method as in claim 4, 5 or 6 wherein said step (c) of processing said digital signal comprises sampling said digital signal into blocks of samples with 0-50% overlap between adjacent blocks.

26. The method as in claim 4, 5 or 6 wherein said step (c) of processing said digital signal comprises sampling said digital signal into blocks from 256 to 2048 samples.

27. The method as in claim 4, 5 or 6 wherein said step (c) of processing said digital signal comprises sampling said digital signal into blocks with 512 samples.

28. The method as in claim 4, 5 or 6 further comprising the step of displaying said machine-readable phonetic symbols after said step (f) of outputting said machine-readable phonetic symbols.

29. The method as in claim 28 further comprising the step of editing said outputted machine-readable phonetic symbols after said step of displaying said machine-readable phonetic symbols.

30. The method as in claim 29 wherein said step of editing said machine-readable phonetic symbols comprises using at least one of a keyboard, a mouse, a digitizing pen, a voice command and an input-positioning device.

31. The method as in claim 28 further comprising the step of editing said word boundaries after said step of displaying said machine-readable phonetic symbols.

32. The method as in claim 4, 5 or 6 wherein said step (i) of visually representing said possible word choice comprises visually representing all the words in any block of speech bracketed by silence intervals at one time.

33. The method as in claim 32 wherein a silence interval is any period of longer than 1/10 of a second in which the ambient noise level does not increase more than 10 dB over the ambient background noise level.

34. The method as in claim 32 wherein a silence interval is any period of longer than 1/15 of a second in

which the ambient noise level does not increase more than 9 dB over the average ambient background noise level.

35. The method as in claim 32 wherein a silence interval is any period between 1/15 of a second and 1/10 of a second in which the ambient noise level does not increase more than about 10 dB over the average ambient background noise level.

36. The method as in claim 4, 5 or 6 further comprising the step of selecting said represented possible word choice.

37. The method as in claim 36 wherein said step of selecting one of said represented possible word choice comprises using at least one of a mouse, keyboard, digitizing pen, voice command and an input-positioning device.

38. The method as in claim 36 wherein said step of identifying word boundaries comprises using adaptive feedback from the step of selecting one of said represented possible word choice as performed on the immediately preceding word.

39. The method as in claim 36 wherein said step of selecting one of said represented possible word choice comprises employing a suitable software/computer combination to make such selection.

40. The method as in claim 36 further comprising audio reproduction of said speech input from said machine-readable phonetic symbols.

41. The method as in claim 6 further comprising the step of selecting said represented possible word choice, wherein said speaker-specific training data is updated via

SDOCID: <WO___9603741A1_IA>

adaptive feedback from said step of selecting said represented possible word choice.

42. The method as in claim 4, 5 or 6 further comprising the step of selecting one of a previously represented possible word choice wherein at least one of said steps (a) through (i) occurs simultaneous with said step of selecting one of a previously represented possible word choice.

43. The method as in claim 4, 5 or 6 further comprising the step of storing said machine-readable phonetic symbols.

44. A method for facilitating speech transcription comprising the steps of:

(a) receiving a human voice speech input as an analog signal;

(b) converting said speech input from said analog signal to a digital signal;

(c) processing said digital signal to extract spectral speech features;

(d) identifying phonemes by comparing said spectral speech features with a reference set of phoneme models and speaker-specific training data;

(e) generating corresponding machine-readable phonetic symbols for said phonemes;

(f) outputting said machine-readable phonetic symbols into a software pre-processor;

(g) displaying said machine-readable phonetic symbols;

(h) editing said machine-readable phonetic symbols;

(i) identifying word boundaries within a substring of said achine-readable phonetic symbols;

(j) editing said word boundaries;

(k) combining said substring into a possible word choice;

(l) ranking said possible word choice; and

(m) visually representing said possible word choice.

45. The method as in claim 44 further comprising the step (n) of selecting one of said represented possible word choices.

46. The method as in claim 45 wherein said steps (a) through (n) are repeated for a word within said human voice speech input and said step (i) of identifying word boundaries includes the use of adaptive feedback from the step (n) of selecting one of said represented possible word choice which occurred for the word preceding said word within said human voice speech input.

47. A method for storing speech as machine-readable phonetic symbols comprising the steps of:

(a) receiving a human voice speech input as an analog signal;

(b) converting said speech input from said analog signal to a digital signal;

(c) processing said digital signal to extract spectral speech features;

(d) identifying phonemes from said spectral speech features; and

(e) generating corresponding machine-readable phonetic symbols for said phonemes.

48. A system for storing speech as machine-readable phonetic symbols comprising:

(a) a device which receives an analog signal representing human speech;

(b) an analog-to-digital convertor in communication with said device for converting said analog signal into a digital signal;

(c) a spectral processor in communication with said analog-to-digital convertor for receiving said digital

signal and extracting spectral speech features from said digital signal; and

(d) a phoneme labeller in communication with said spectral processor for receiving said spectral speech features and for identifying phonemes from said spectral speech features.

49. The system as in claim 48 wherein said analog signal representing human speech is an electromagnetic signal.

50. A system for storing speech as machine-readable phonetic symbols comprising:

(a) a device which receives a digital signal representing human speech;

(b) a spectral processor in communication with said device for receiving said digital signal and extracting spectral speech features from said digital signal;

(c) a phoneme labeller in communication with said spectral processor for receiving said spectral speech features, for identifying phonemes from said spectral speech features; and for generating corresponding machine-readable phonetic symbols for said phonemes; and

(d) a storage device for storing said machine-readable phonetic symbols.

51. The system as in claim 50 wherein said digital signal representing speech is an electromagnetic signal.

52. A method of speech reproduction comprising re-synthesis of stored machine-readable phonetic symbols which represent a compressed form of speech.

53. The method as in claim 4 further comprising the step of ranking said possible word choice in order of the most recent usage, wherein the word choice with the least

number of words since the last usage of said word choice is ranked first.

54. The method as in claim 53 wherein said step (i) of visually representing said possible word choice comprises displaying said possible word choice in order of the most recent usage, wherein the word choice with the least number of words since the last usage of said word choice is ranked first.

55. The method as in claim 4 further comprising the step of ranking said possible word choice in order of frequency used, wherein the word previously used most often is ranked first.

56. The method as in claim 55 wherein said step (i) of visually representing said possible word choice comprises displaying said possible word choice in order of frequency used, wherein the word previously used most often is ranked first.

57. The method as in claim 4 further comprising the step of ranking said possible word choice by the time elapsed since said possible word choice was last used, wherein the word choice temporally closest to its last usage is ranked first.

58. The method as in claim 57 wherein said step (i) of visually representing said possible word choice comprises displaying said possible word choice by the time elapsed since said possible word choice was last used, wherein the word choice temporally closest to its last usage is ranked first.

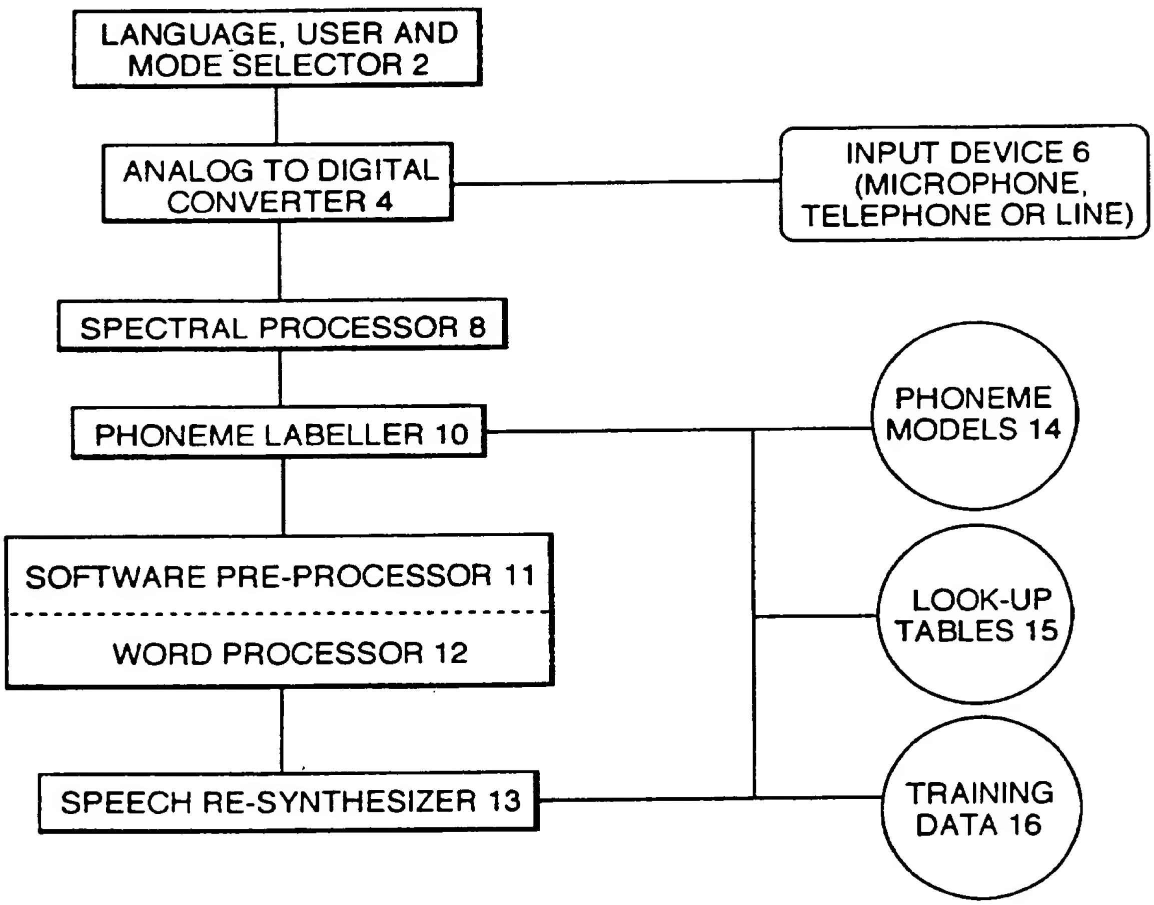


FIG. 1.

SDOCID: <WO___9603741A1_IA>

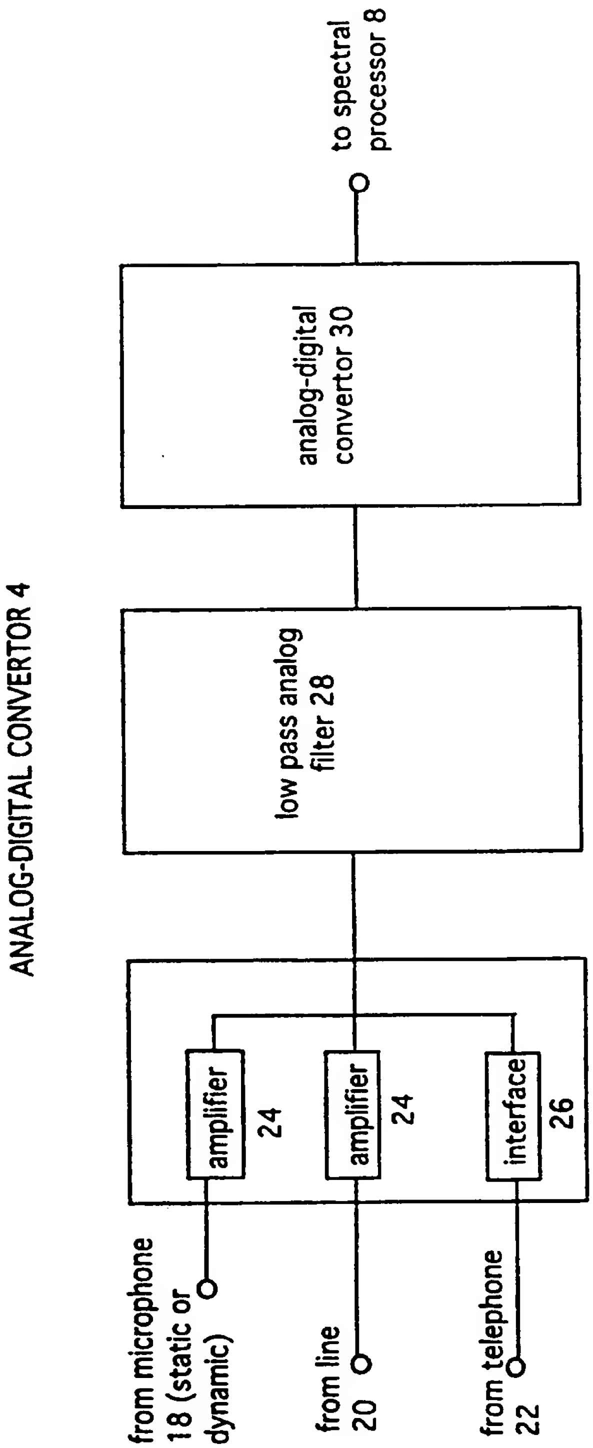

ANALOG-DIGITAL CONVERTOR 4
from microphone 18 (static or dynamic)
from line 20
from telephone 22
amplifier 24
amplifier 24
interface 26
low pass analog filter 28
analog-digital convertor 30
to spectral processor 8

FIGURE 2

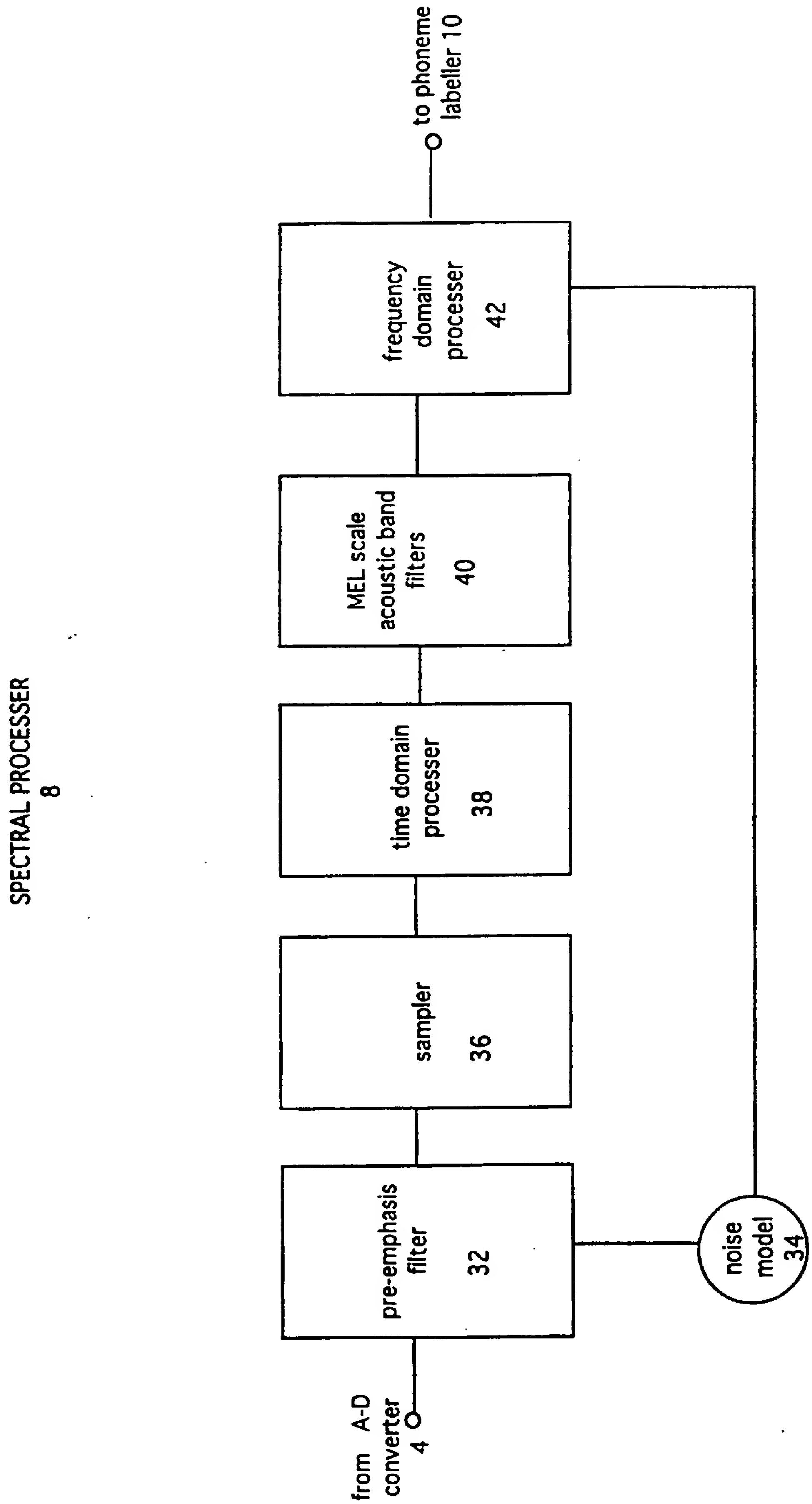

SPECTRAL PROCESSER
8
from A-D converter 4
pre-emphasis filter
32
sampler
36
time domain processer
38
MEL scale acoustic band filters
40
frequency domain processer
42
to phoneme labeller 10
noise model 34

FIGURE 3

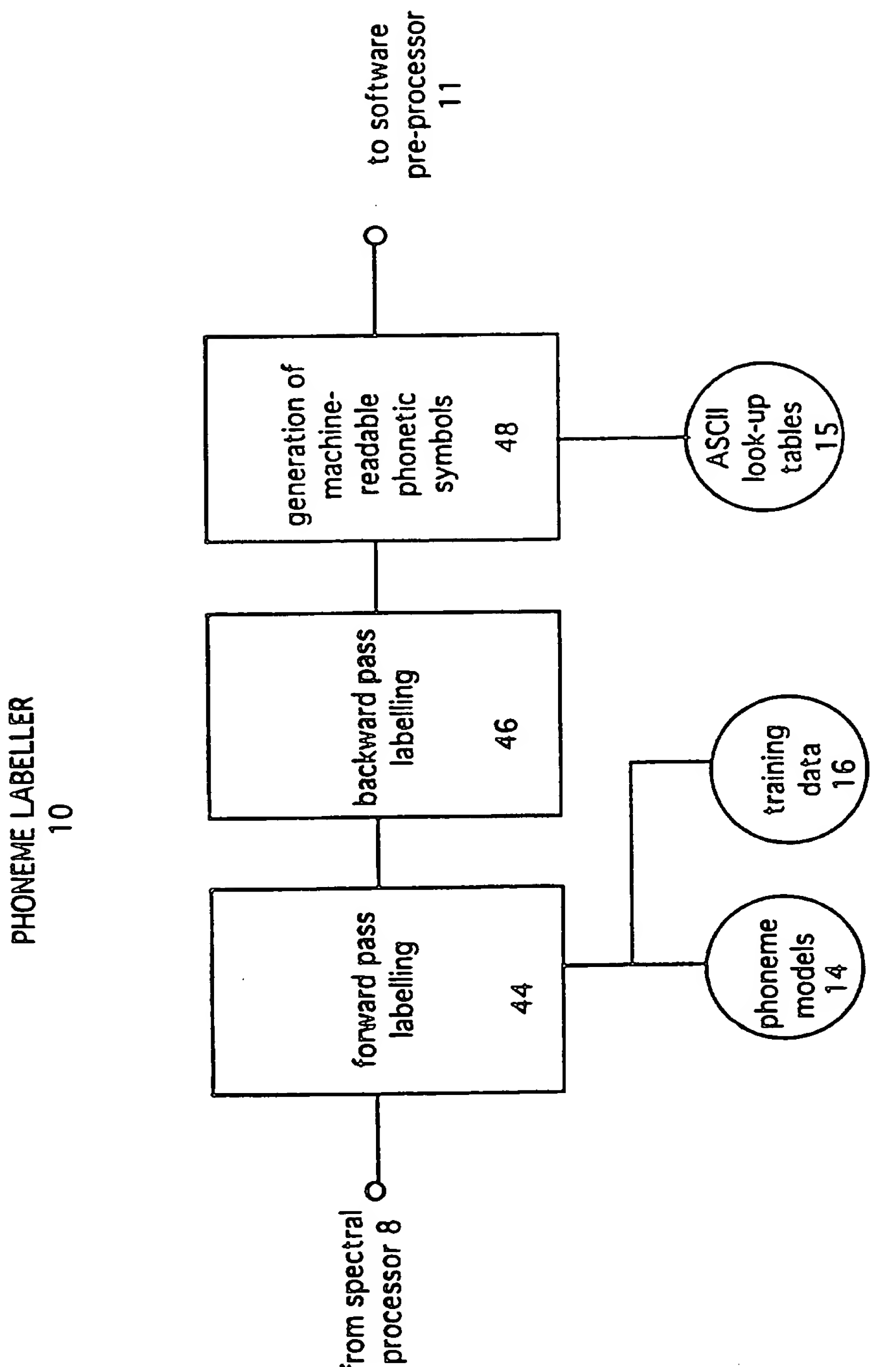

PHONEME LABELLER
10
from spectral
processor 8
forward pass
labelling
44
backward pass
labelling
46
generation of
machine-
readable
phonetic
symbols
48
to software
pre-processor
11
phoneme
models
14
training
data
16
ASCII
look-up
tables
15

FIGURE 4

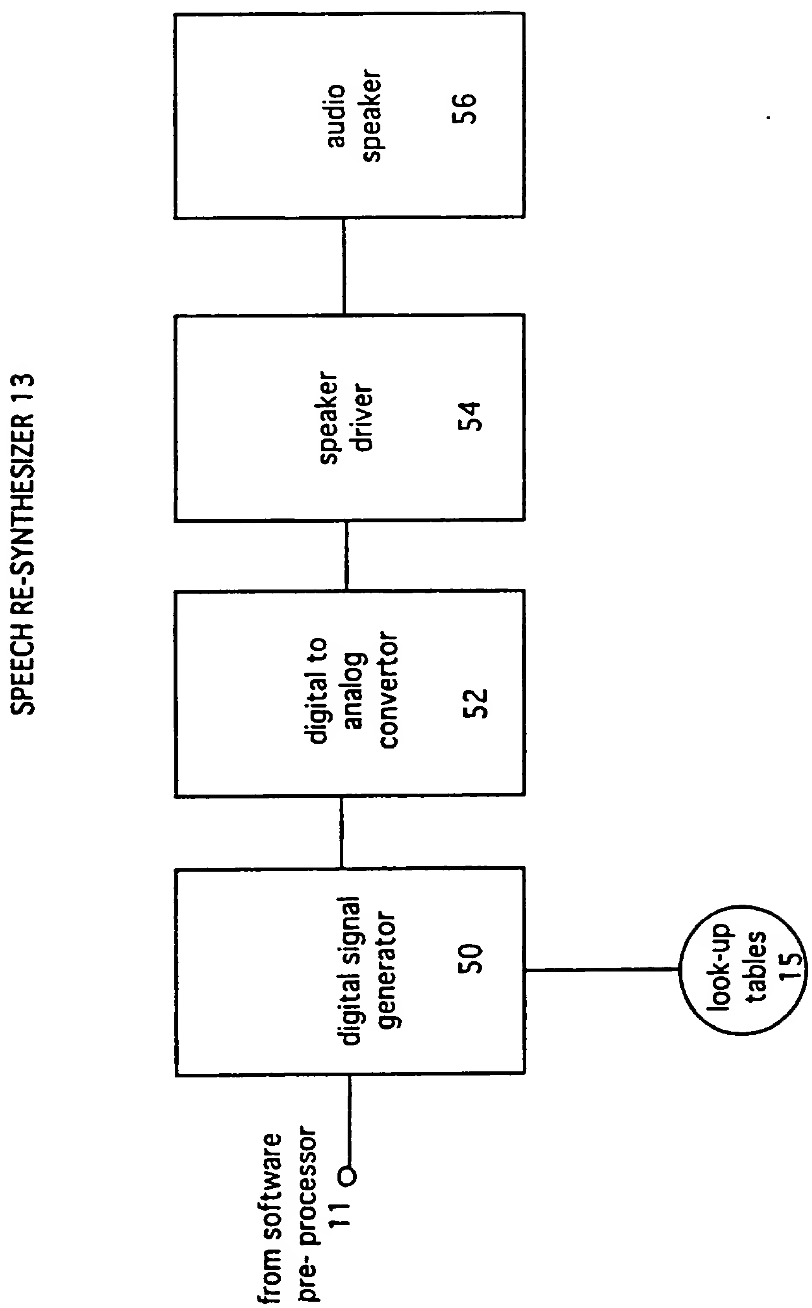

SPEECH RE-SYNTHESIZER 13
from software pre- processor 11
digital signal generator 50
look-up tables 15
digital to analog convertor 52
speaker driver 54
audio speaker 56

FIGURE 5

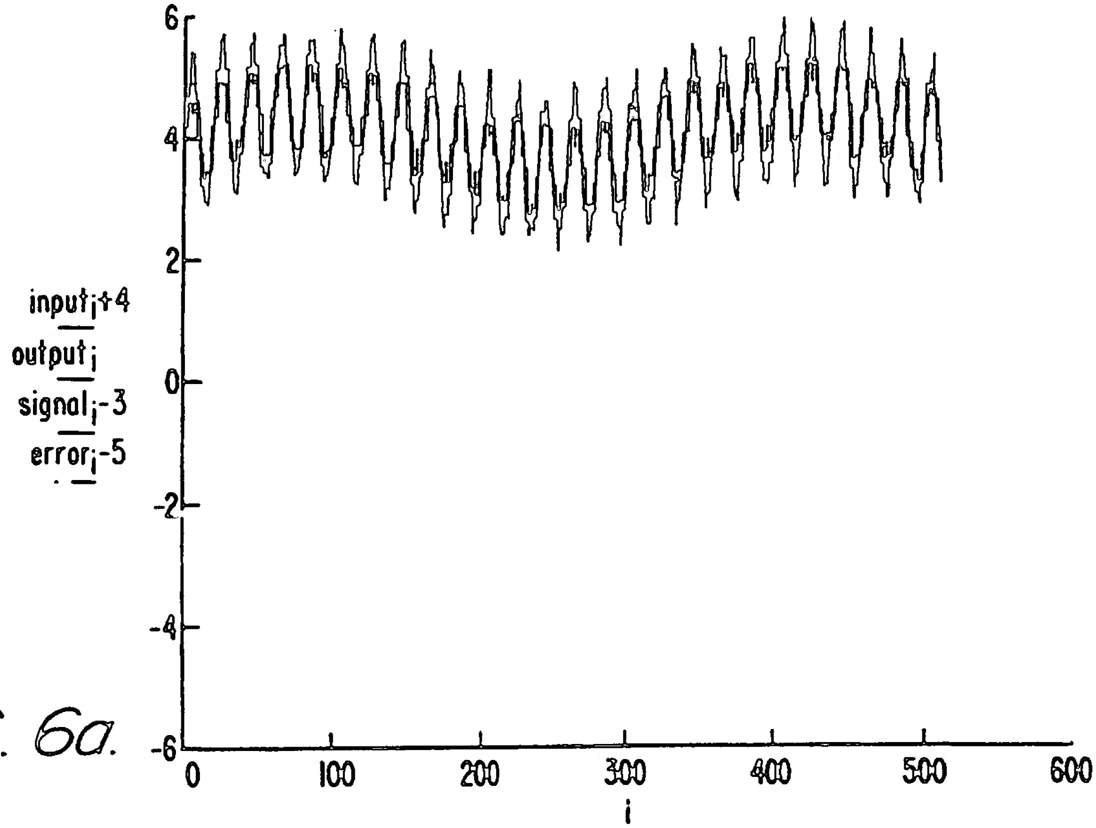


FIG. 6a.

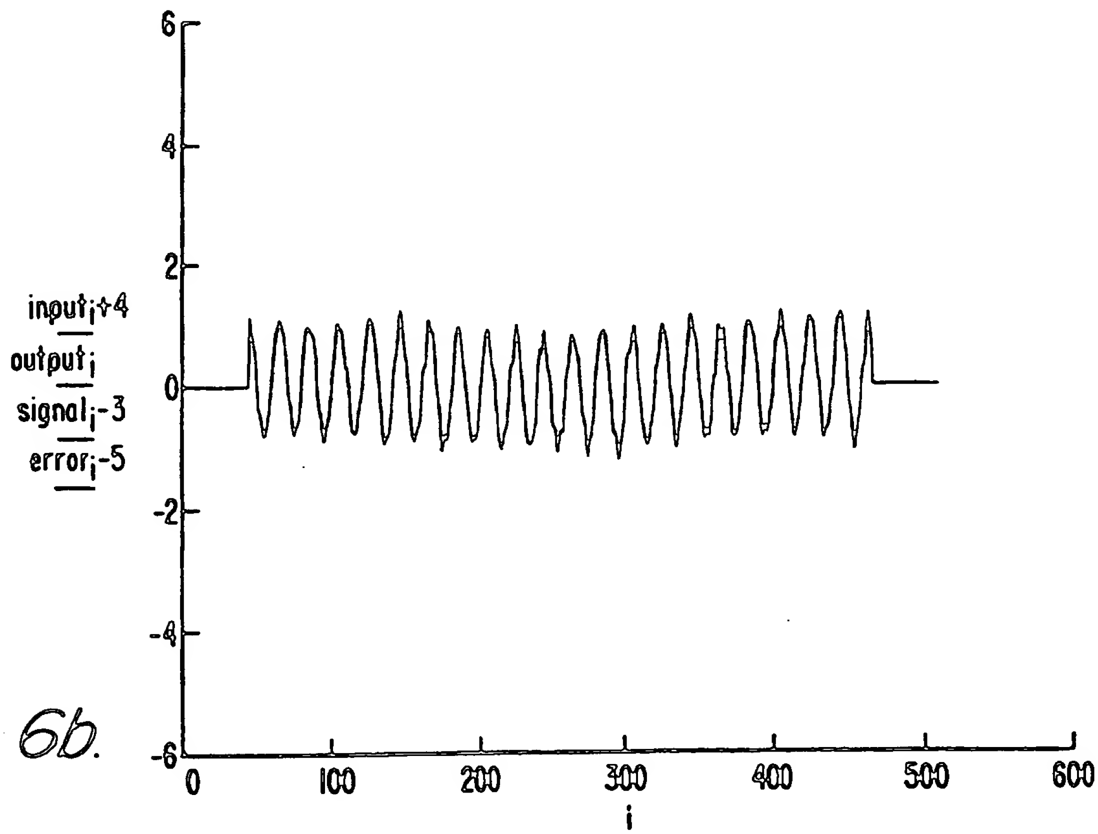


FIG. 6b.

SDOCID: <WO___9603741A1_IA>

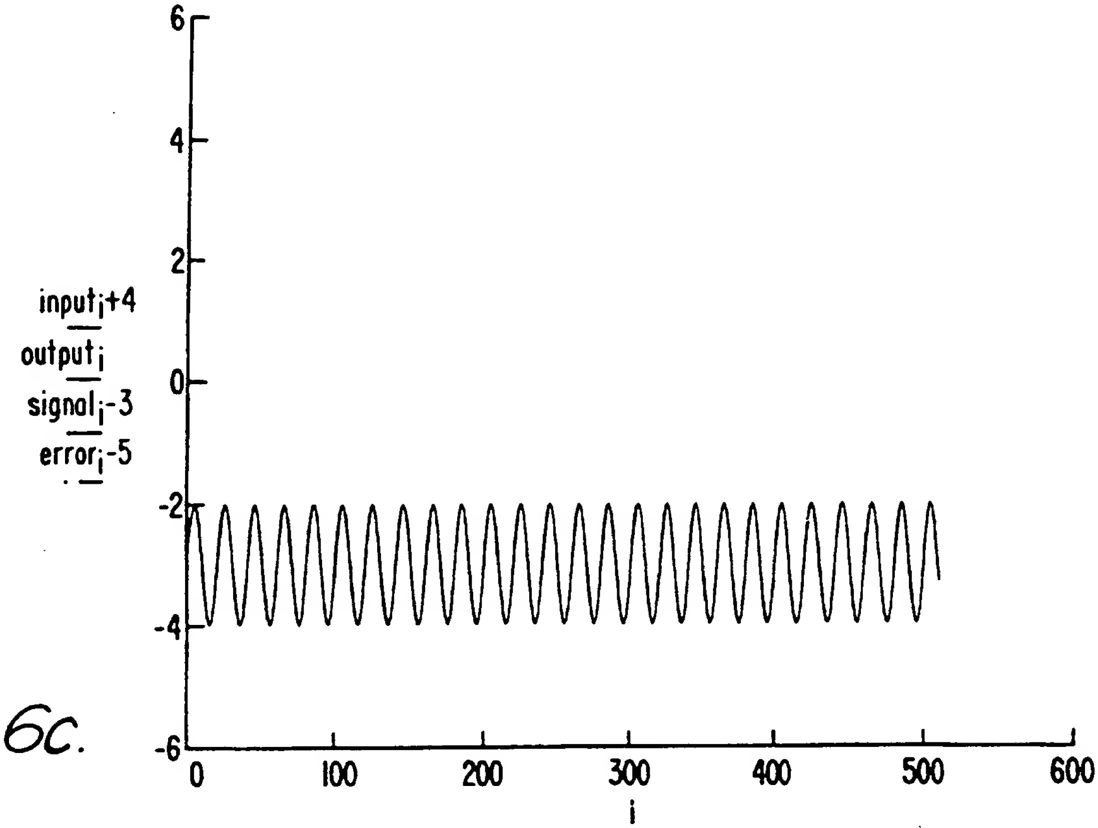


FIG. 6c.

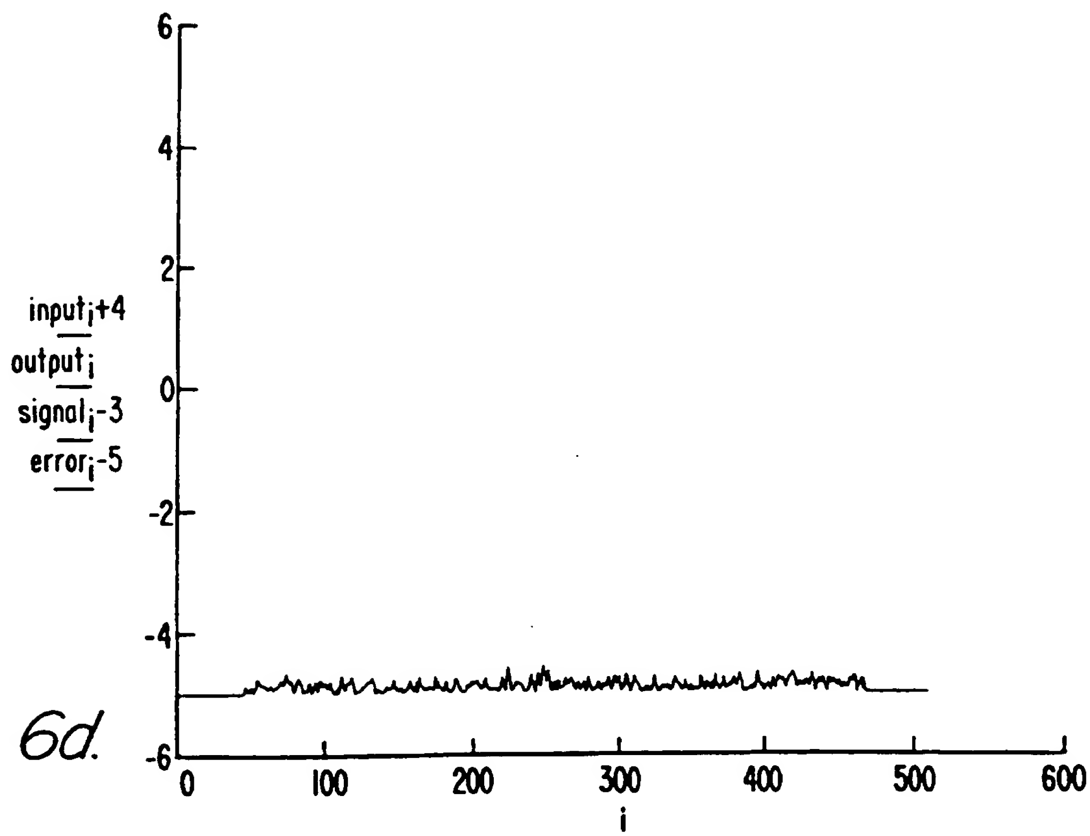


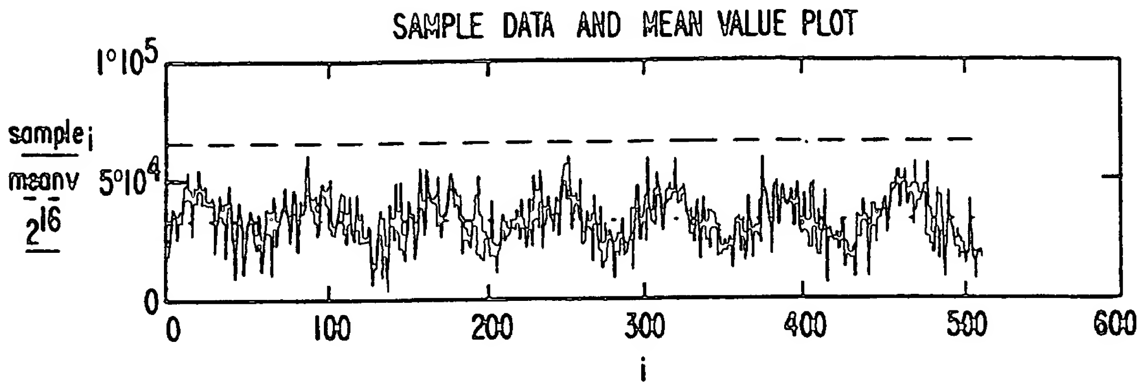
FIG. 6d.

SDOCID: <WO___9603741A1_IA>

ns :=511

i :=0..ns

$$mean := \frac{\sum_{k=0}^{ns} sample_k}{ns+1} \qquad mean = 3.298 \cdot 10^4$$ COMPUTED ZERO REFERENCE



range :=2^{16}

$$signal_i := \frac{sample_i - meanv}{range}$$ ESTIMATED NORMALIZED SIGNAL

FIG. 7a.

ABSOLUTE POWER MODEL

$$power1 := \frac{\sum_{k=0}^{ns} |signal_k|}{ns+1}$$

power1 =0.142

FIG. 7b.

SQUARED POWER MODEL

$$power2 := \sqrt{\frac{\sum_{k=0}^{ns} (signal_k)^2}{ns+1}}$$

power2 =0.172

FIG. 7c.

SDOCID: <WO___9603741A1_IA>

FIG. 8a.

ESTIMATED SIGNAL PLOT

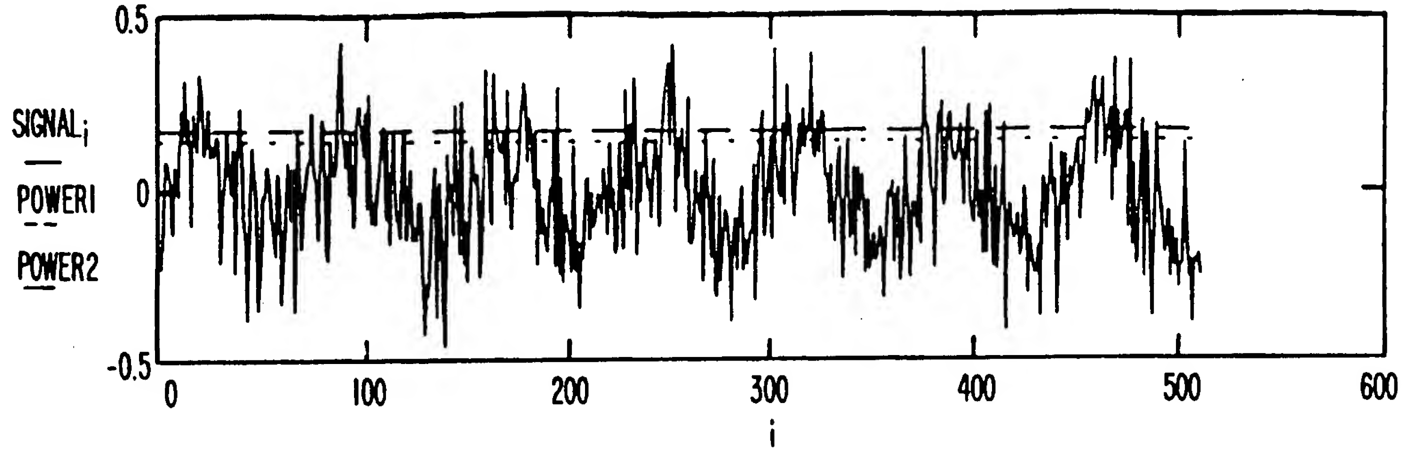


FIG. 8b.

$$\mathrm{sign}(i) := \mathrm{if}(\mathrm{signal}_i \cdot \mathrm{signal}_{i+1} > 0, 0, 1)$$

$$\mathrm{zeros} := \frac{\sum_{k=0}^{ns-1} \mathrm{if}(\mathrm{sign}(k), 0, 1)}{2} \qquad \mathrm{zeros} = 154.5 \quad \text{ZERO CROSSING FREQUENCY}$$

.SDOCID: <WO___9603741A1_IA>

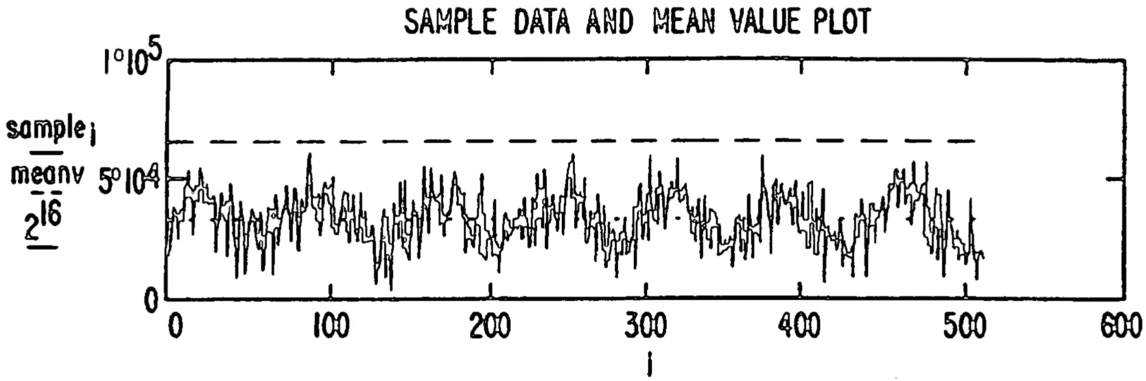


FIG. 9a.

$$dt := 0.00005$$

$$j := 0..ns \qquad freq_j := \frac{j}{dt \cdot ns} \qquad win_j := 0.54 - 0.46 \cdot \cos\left(2 \cdot \pi \cdot \frac{j}{ns}\right)$$

$$signal_j := sample_j \cdot win_j$$

FIG. 9b.

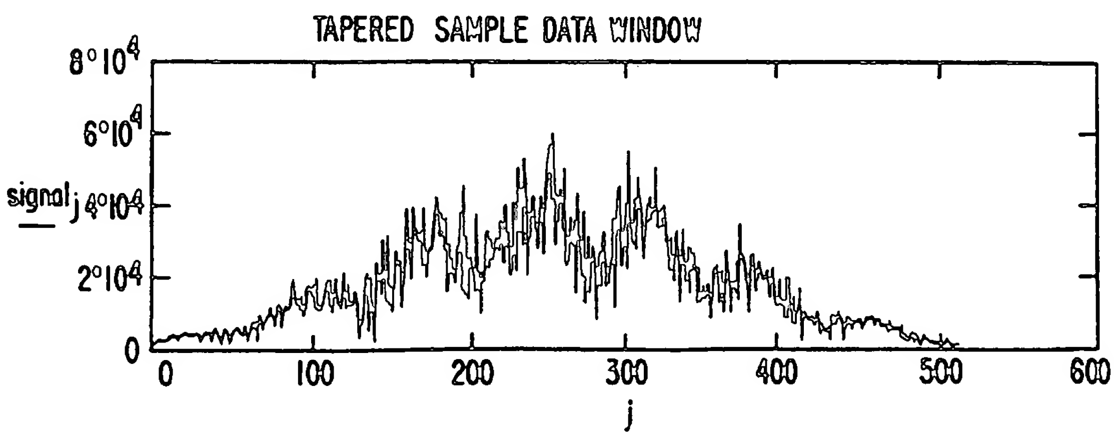


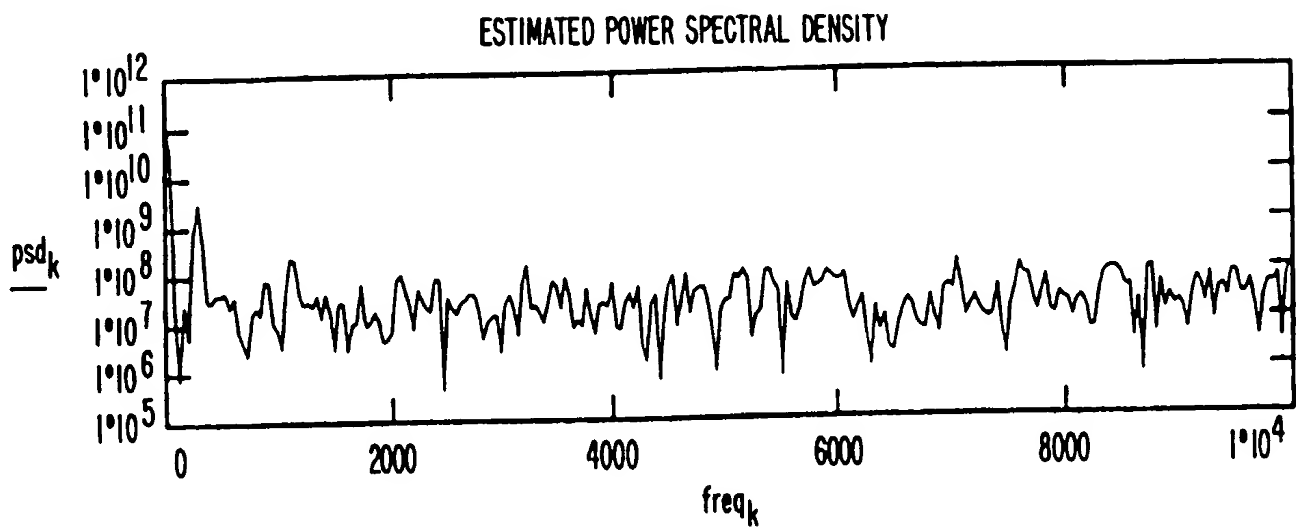
FIG. 9c.

SDOCID: <WO___9603741A1_IA>

$$psd := \overrightarrow{\left[\left(\left|fft(signal)\right|\right)^2\right]}$$

$k := 0 .. \frac{ns}{2}$ $\quad lpsd_k := \log(psd_k)$

FIG. 9d.

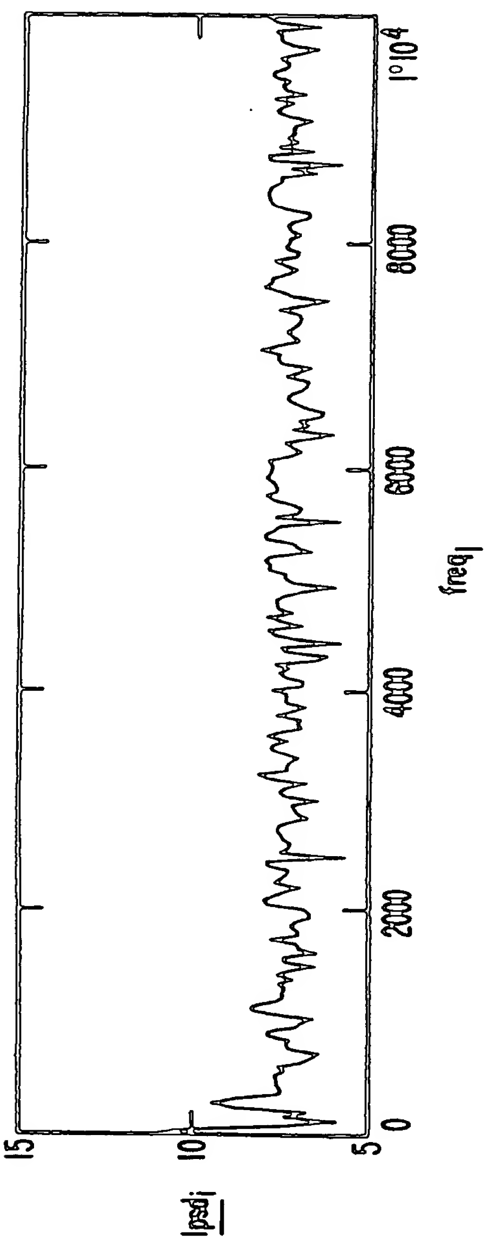


FORMANTS $fdata_0 = 270$ $\quad fdata_1 = 2.29 \cdot 10^3$ $\quad fdata_2 = 3.07 \cdot 10^3$

$1 := 0 .. \frac{ns}{2}$

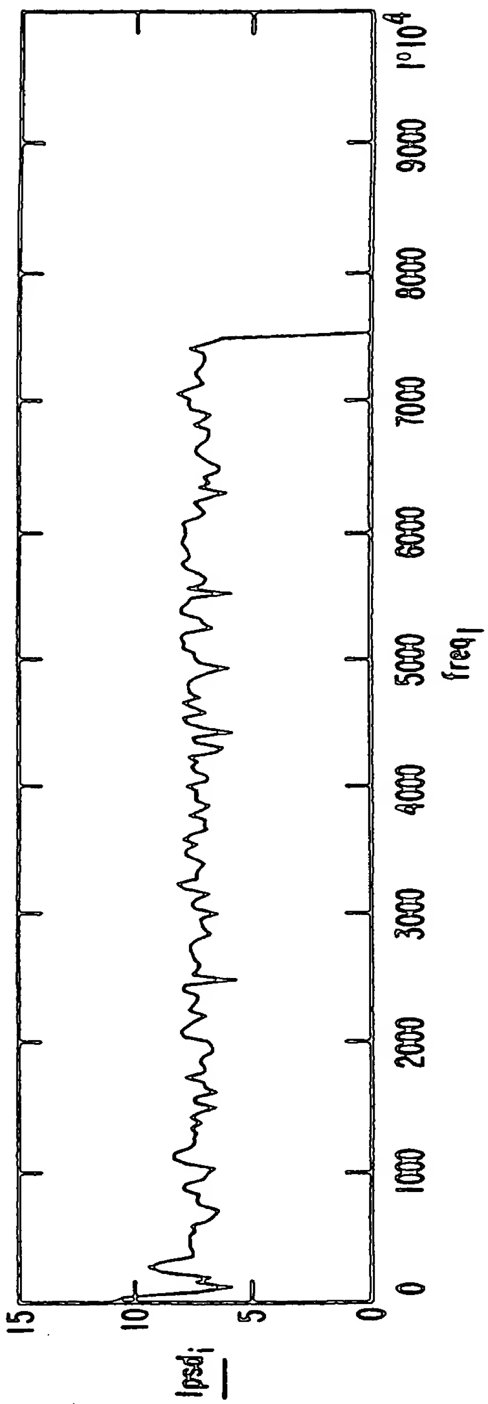
FIG. 9e.

SDOCID: <WO___9603741A1_IA>



m :=192..ns   lpsd$_m$:0

FIG. 9f.



m :=0..127 len :=length(lpsd) len=512 ns=511

$xpsd := \left[\overrightarrow{\left(|fft(lpsd)| \right)^2} \right]$

FIG. 9g.

SDOCID: <WO___9603741A1_IA>

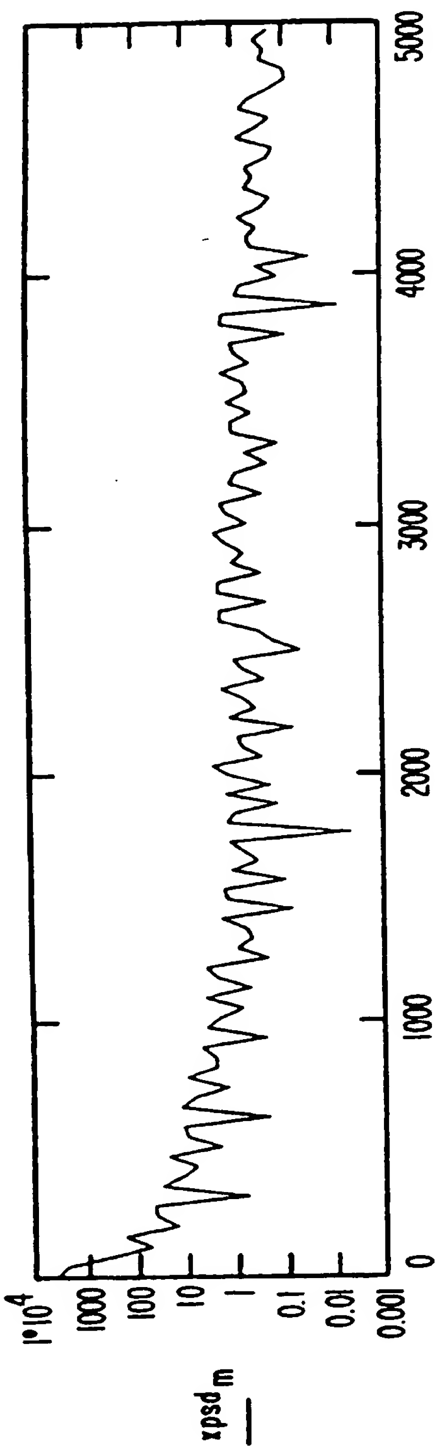


FIG. 9h.

nFeat :=3 NUMBER OF FEATURES TO COMBINE

nPhon :=5 NUMBER OF PHONEME ALTERNATIVES

nRes :=10 DISCRETE RESOLUTIONS OF MEMBERSHIP FUNCTION

SAMPLE DATA FEATURE VALUES $f := \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}$

AVERAGE OF FEATURE VALUES fc :=mean(f) fc=2

PHONEME FEATURE VALUES $p := \begin{pmatrix} 2 & 2 & 1 & 1 & 3 \\ 1 & 3 & 2 & 3 & 2 \\ 3 & 2 & 4 & 1 & 2 \end{pmatrix}$

AVERAGE OF PHONEME VALUES i :=0..nPhon-1

$pc_i := \mathrm{mean}(p^{\langle i \rangle})$ $pc = \begin{bmatrix} 2 \\ 2.333 \\ 2.333 \\ 1.667 \\ 2.333 \end{bmatrix}$

COMPUTE JOINT MEMBERSHIP FUNCTION FOR EACH PHONEME

del :=1 eps :=0.0

uu_i :=wx(fc-del,fc-eps,fc+eps,fc+del,pc_i-del,pc_i-eps,pc_i+eps,pc_i+del,1,1)

POSSIBILITIES FOR EACH PHONEME $uu = \begin{bmatrix} 0.5 \\ 0.36 \\ 0.36 \\ 0.5 \\ 0.36 \end{bmatrix}$

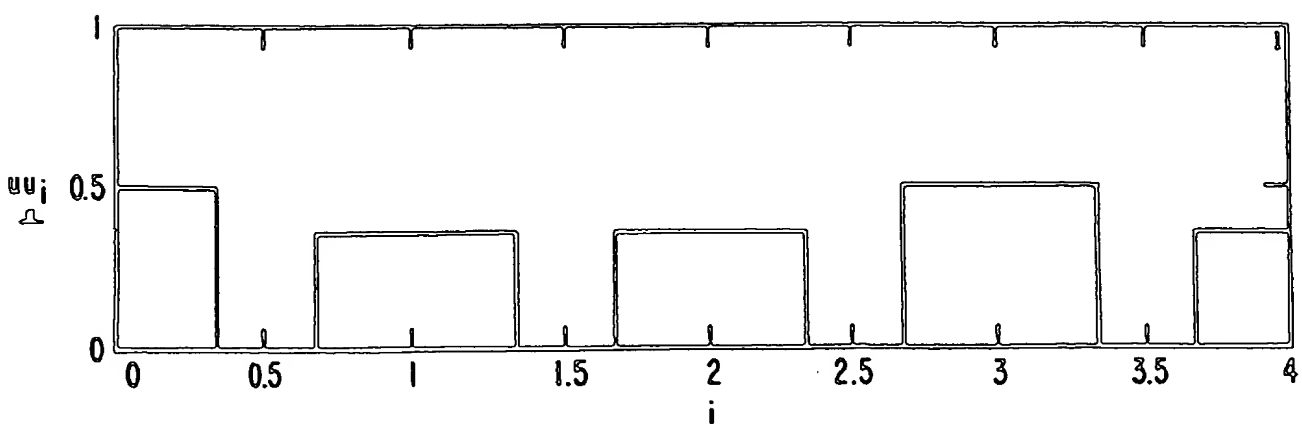


FIG. 10a.

SDOCID: <WO___9603741A1_IA>

FIG. 10b.

COMPUTE VALUES OF PHONEME POSSIBILITY FUNCTIONS FOR DISPLAY

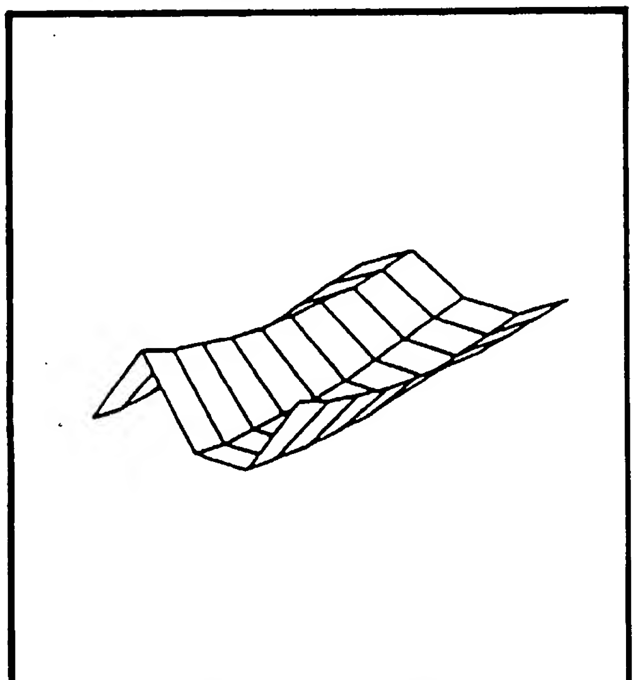
$k := 0..nRes-1$

$$u_{i,k} := wx\left(fc-1, fc, fc, fc+1, pc_i-1, pc_i, pc_i, pc_i+1, 1, 1, \frac{k}{nRes}\right)$$

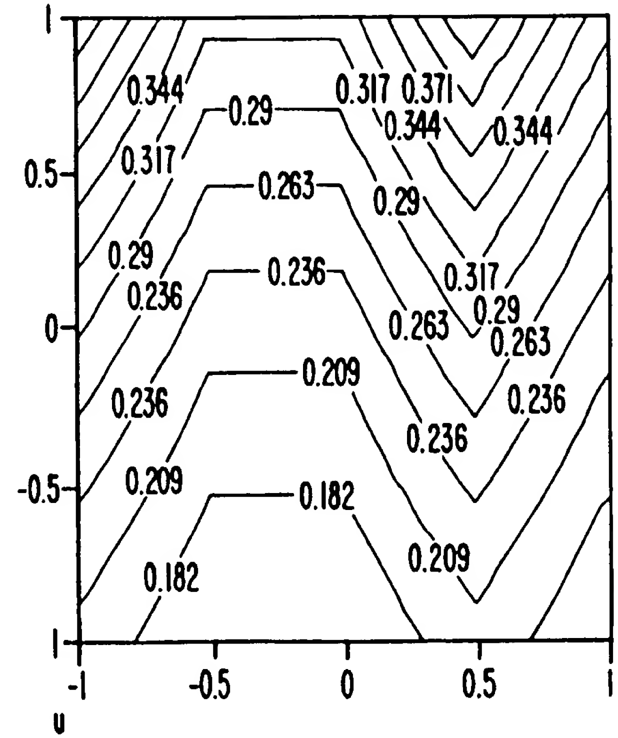
VALUES OF JOINT MEMBERSHIP FUNCTION

$$u=\begin{bmatrix} 0.2 & 0.217 & 0.236 & 0.257 & 0.281 & 0.308 & 0.338 & 0.372 & 0.41 & 0.452 \\ 0.155 & 0.167 & 0.18 & 0.195 & 0.211 & 0.229 & 0.25 & 0.273 & 0.298 & 0.327 \\ 0.155 & 0.167 & 0.18 & 0.195 & 0.211 & 0.229 & 0.25 & 0.273 & 0.298 & 0.327 \\ 0.2 & 0.217 & 0.236 & 0.257 & 0.281 & 0.308 & 0.338 & 0.372 & 0.41 & 0.452 \\ 0.155 & 0.167 & 0.18 & 0.195 & 0.211 & 0.229 & 0.25 & 0.273 & 0.298 & 0.327 \end{bmatrix}$$

SURFACE PLOT OF MEMBERSHIP FUNCTIONS



CONTOUR PLOT OF MEMBERSHIP VALUES



DOCID: <WO___9603741A1_IA>

NUMBER OF PHONEMES $np := 5$

NUMBER OF SAMPLE DATA BLOCKS $ns := 9$

POSSIBILITY MATRIX FOR SAMPLE BLOCKS

$$pm := \begin{bmatrix} .4 & .36 & .4 & .14 & .06 & .06 & .05 & .04 & .04 & .04 \\ .11 & .09 & .15 & .35 & .4 & .36 & .31 & .1 & .11 & .11 \\ .04 & .06 & .06 & .11 & .04 & .05 & .15 & .35 & .36 & .4 \\ .3 & .34 & .3 & .09 & .11 & .15 & .11 & .06 & .06 & .06 \\ .09 & .11 & .05 & .16 & .3 & .34 & .29 & .15 & .09 & .09 \\ .06 & .04 & .04 & .14 & .09 & .04 & .09 & .3 & .34 & .3 \end{bmatrix}$$

DETERMINE THE TOP TWO RANKED PHONEME ASSIGNMENTS OF EACH SAMPLE BLOCK

$i := 0..ns \qquad pmax_i := \max\left(pm^{<i>}\right)$

$j := 0..np \qquad parg_{j,i} := \dfrac{pm_{j,i}}{pmax_i} \qquad pass_{j,i} := if\left(parg_{j,i} < 1.0, 0, 1\right)$

$pm_{j,i} := if\left(pass_{j,i} > 0, 0, pm_{j,i}\right)$

$i := 0..ns \qquad pmax_i := \max\left(pm^{<i>}\right)$

$j := 0..np \qquad parg_{j,i} := \dfrac{pm_{j,i}}{pmax_i} \qquad pass_{j,i} := if\left[\left(parg_{j,i} < 1.0\right), pass_{j,i}, 2\right]$

THE ASSIGNMENT MATRIX FOR THE TOP TWO CHOICES IN THE SET OF SAMPLE BLOCKS IS

$$pass = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 & 2 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 \end{bmatrix}$$

FIG. 11.

SDOCID: <WO___9603741A1_IA>

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 95/09130

A. CLASSIFICATION OF SUBJECT MATTER

G 10 L 5/06, G 10 L 7/08, G 10 L 9/06, G 10 L 9/18

According to International Patent Classification (IPC) or to both national classification and IPC 6

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G 10 L, H 04 M

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP, A, 0 248 593 (SPEECH SYSTEMS, INC.) 09 December 1987 (09.12.87), fig. 1; abstract; claims 1-30. --	1-10, 44, 47-51
X, P	EP, A, 0 645 757 (XEROX CORP.) 29 March 1995 (29.03.95), fig. 1; abstract; claims 1-12. --	1-58
A	US, A, 5 289 523 (VASILE et al.) 22 February 1994 (22.02.94), fig. 3; abstract; claims 1-21. ----	1-58

☐ Further documents are listed in the continuation of box C.

☐ Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 October 1995	10. 11. 95

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax (+31-70) 340-3016	BERGER e.h.

ANHANG	ANNEX	ANNEXE
zum internationalen Recherchenbericht über die internationale Patentanmeldung Nr.	to the International Search Report to the International Patent Application No.	au rapport de recherche international relatif à la demande de brevet international n°

PCT/US 95/09130 SAE 114738

In diesem Anhang sind die Mitglieder der Patentfamilien der im obengenannten internationalen Recherchenbericht angeführten Patentdokumente angegeben. Diese Angaben dienen nur zur Unterrichtung und erfolgen ohne Gewähr.

This Annex lists the patent family members relating to the patent documents cited in the above-mentioned international search report. The Office is in no way liable for these particulars which are given merely for the purpose of information.

La presénte annexe indique les membres de la famille de brevets relatifs aux documents de brevets cités dans le rapport de recherche international visée ci-dessus. Les reseignements fournis sont donnés à titre indicatif et n'engagent pas la responsibilité de l'Office.

Im Recherchenbericht angeführtes Patentdokument Patent document cited in search report Document de brevet cité dans le rapport de recherche	Datum der Veröffentlichung Publication date Date de publication	Mitglied(er) der Patentfamilie Patent family member(s) Membre(s) de la famille de brevets	Datum der Veröffentlichung Publication date Date de publication
EP A1 248593	09-12-87	JP A2 63066600	25-03-88
		US A 5054085	01-10-91
		AT E 45831	15-09-89
		AU A1 28312/84	21-11-85
		AU B2 580000	22-12-88
		CA A1 1216944	20-01-87
		DE CO 3479542	28-09-89
		EP A1 126420	28-11-84
		EP B1 126420	23-08-89
		JP A2 59221198	12-12-84
		US A 4718096	05-01-88
EP A1 645757	29-03-95	JP A2 7175497	14-07-95
US A 5289523	22-02-94	keine - none - rien	

SDOCID: <WO___9603741A1_IA>